

From Feeds to Inboxes: A Comparative Study of Polarization in Facebook and Email News Sharing

Hema Yoganarasimhan * Irina Yakovetskaya
University of Washington *Stanford University*

December 17, 2023

Abstract

This study explores the polarization of news content shared on Facebook compared to email, using data from the New York Times' Most Emailed and Most Shared lists over 2.5 years. Employing Latent Dirichlet Allocation (LDA) and Large Language Models (LLMs), we find that highly polarized articles are more likely to be shared on Facebook (vs. email), even after accounting for factors like topics, emotion, and article age. Additionally, distinct topic preferences emerge, with social issues dominating Facebook shares and lifestyle topics prevalent in emails. Contrary to expectations, political polarization of articles shared on Facebook did not escalate post-2020 election. We introduce a novel approach to measuring polarization of text content that leverages generative AI models like ChatGPT, which is both scalable and cost-effective. This research contributes to the evolving intersection of Large Language Models (LLMs), social media, and polarization studies, shedding light on descriptive patterns of content dissemination across different digital channels.

Keywords: Polarization, Social Media, Large Language Models (LLMs), Facebook, News

*Please address all correspondence to: hemay@uw.edu and irinay@stanford.edu. We would like to thank Mark Forehand, Ali Goli, Simha Mummalaneni, and Omid Rafeian for feedback that has significantly improved the paper.

1 Introduction

Over the last two decades, we have seen significant shifts in how people consume and share news. Historically, users obtained news directly from established news sources by reading newspapers or news websites. However, over the years, the way users are exposed to news and consume news has changed because of the availability of digital communication formats that allow users to share news articles with their peers (e.g., emails and messages). More recently, with the rise of social media platforms such as Facebook and Twitter, users share news from news websites on their social media pages, which in turn can be re-shared by others. Thus, users can now obtain news without going to the original news website. While these changes have led to easier access to news, there have been concerns that these changes have skewed news sharing and consumption toward polarizing topics.

Earlier research has provided some support for the idea that social media platforms promote polarized discourse through mechanisms such as homophilic network formation, echo chambers, and filter bubbles, and suggested that personalization algorithms can amplify these effects (Shin and Kadiyala, 2022; Allcott et al., 2020; Barberá et al., 2015; Cinelli et al., 2021; Levy, 2021; Pariser, 2011; Persily and Tucker, 2020; Sunstein, 2018). On the other hand, some emerging research has questioned this narrative (Bail et al., 2018; Boxell et al., 2017; Eady et al., 2019; Gentzkow and Shapiro, 2011). For instance, some early work shows that users’ exposure to political news on Facebook is primarily driven by the content shared by their friends rather than the algorithm used to determine users’ news feeds (Bakshy et al., 2015).

In this paper, we examine one specific aspect of this issue – is the content seeded on social media (Facebook) systematically more polarized than that shared via more personal channels (email) after controlling for a variety of article features such as the content and topics in the article?

To answer this question, we use the data from the New York Times (NYTimes) Most Emailed and Most Shared in this study. Our analysis consists of four main steps. First, for our data collection efforts, we focus on the Trending section of NYTimes, which lists the top 20 most emailed and the top 20 most shared articles on Facebook over the last 24 hours. Both these lists are based on the data gathered by the share buttons on each NYTimes article. We collect daily data on these rank-ordered lists for a ≈ 2.5 -year period starting from January 1, 2019, to May 30, 2021. In addition, we collect the metadata and text for each article that appears at least once across the two lists (a total of 13508 unique articles). In the second step, we use a Latent Dirichlet Allocation (LDA) model to recover the distribution of topics in each article based on the unstructured text in the entire corpus. Third, we obtain measures of polarization for all the articles in our corpus using the newly developed Large Language Models (LLMs). We validate these LLM-based measures of polarization using a user survey. Finally, we quantify the relationship between a topic’s prevalence in an article and its relative popularity on Facebook vs. email using a descriptive model. In the process, we control for time-varying article-specific shocks that can affect an article’s ranking on both lists.

We now discuss our main findings. We find that more polarizing articles have a higher relative likelihood of being shared on Facebook compared to email, even after controlling for a series of confounding factors such as the topics covered in the article, news section, emotional content, and age of the article. In addition,

we find that socially relevant and general interest topics are more likely to be posted on social media compared to being shared via email (after controlling for the polarization score and other control variables). Specifically, articles on topics such as *Books, Business, Animals, Food, Real-estate, Nature, and Health Research and Lifestyle Advice* are more commonly sent through email. In contrast, articles on topics such as *Election investigations, Covid vaccine, Russia, Women’s Issues and Sexual Harassment, Coronavirus Pandemic, Black Lives Matter* etc., are more likely to be posted on social media. Next, we examine if the articles seeded on Facebook become more politically polarized over time. This question is motivated by the discussions around the exacerbation of polarization on social media platforms after the 2020 election (Jurkowitz et al., 2020). Interestingly, we find the opposite – polarization scores of articles play a smaller role in predicting the differences across the two lists after the election.

Our paper contributes to the literature on social media and polarization in two ways. First, we show that the recently developed generative AI models such as ChatGPT can be used to measure the polarization of text content in a scalable low-cost fashion (something not feasible with user surveys). While some recent research has shown that these models produce responses consistent with true user preferences in certain settings (Brand et al., 2023), others have shown the opposite (Goli and Singh, 2023), i.e., there is no consensus on which types of user preferences these models can accurately simulate. Our findings provide some initial evidence in support of their accuracy in characterizing polarization and political bias. We hope this will help spur further research on this topic. Second, we show that the content seeded on social media websites is systematically different from that shared through other media formats, both in terms of polarization as well as the distribution of topics covered. While the exact causes of these descriptive findings are hard to pin down with our data (we discuss some non-exhaustive set of possibilities in §3.3.2), they nevertheless suggest that social media websites tend to attract more polarized content even at the content seeding stage, even before the explicit influence on recommendation algorithms on social media.

2 Data

Our data comes from the Trending section in the NYTimes website, which has two rank-ordered lists – (1) Most Emailed and (2) Popular on Facebook; henceforth referred to as M-Emailed and M-Facebook for convenience. Figure 1 illustrates an example of the two lists. The two lists are constructed based on the data from the share buttons on each article, and then rank-ordered accordingly. To be able to share articles (on email or Facebook) using the share buttons, the user needs to be logged in.¹

We obtained data on these two lists using Internet Archive for a ≈ 2.5 -year period, starting from January 1, 2019, to May 30, 2021.² We parse this data for each day for the time closest to noon. Data are missing at random for some days, and we have data on a total of 697 days in our observation period. Over this period, we see 13688 unique articles across both lists. In addition, we used Article Search API to retrieve article

¹Re-shares of Facebook posts containing NYTimes articles or direct link sharing on Facebook/email are not counted when generating these ranks since those are internal to Facebook. Thus, these lists represent the sharing behavior of logged-in NYTimes users.

²Internet Archive has data from late 2015, but the NYTimes Trending section was in beta till 2017. Data from 2018 exists but has numerous missing observations, and is therefore not reliable. Hence, we focus our data collection efforts from 2019 onwards.

Most Emailed

Most emailed articles today



View Photo Gallery for The New York Times

How a Butterfly Refuge at the Texas Border Became the Target of Far-Right Lies

Climate Change Enters the Therapy Room



Mel Mermelstein, Holocaust Survivor Who Sued Deniers, Dies at 95



Cremation Borrows a Page From



(a) M-Emailed list

Popular on Facebook

Most shared articles on Facebook today



Andrew Harnly/Associated Press

Overhaul of Electoral Count Act 'Absolutely' Will Pass, Manchin Says

Effort to Rescue a 5-Year-Old Transfixes Morocco, Only to End Sadly



Three Dead and One Wounded in Shooting Near Milwaukee, Police Say



Queen Elizabeth Paves the Way for



(b) M-Facebook list

Figure 1: Snapshot of NYTimes trending lists

metadata such as headline, publication date, abstract, and section (Dev Portal, 2022).³ See Table A1 in the Appendix for details. Of the 13688 articles, 13508 were accessible and had both metadata and full text available.⁴

We now conduct a preliminary analysis of the similarities and differences between the two lists. First, we pool all the articles over the entire observation period and examine the overlap between the two lists. There is only a small amount of overlap between the two lists – only 2884 of the 13688 articles (i.e., 20% of articles) appear on both lists at least once. Among the rest of the articles, 6440 articles appear only in the M-Emailed list, while 4404 articles appear only in the M-Facebook list. Next, we examine the overlap in the two lists on any given day. On average, 5.5 articles (out of 20) appear on both lists on any given day, i.e., over 14 articles are different across the two lists at any given point in time see Figure A1 in Appendix §A.1). Thus, there are significant differences in the articles in the two lists (both across days and on any given day).

Next, we examine whether the articles that appear on these two lists are systematically different on basic attributes derived from metadata. We find that articles shared on Facebook, on average, are shorter with a mean of 1390 words compared to articles in the M-Emailed list, which average 1540 words (a t-test confirmed that this difference is significant).⁵ Further, in the M-Emailed list, over 30% come from the Opinion section,

³NYTimes sometimes changes the headlines of news articles. However, this does not affect our analysis since we work with individual article IDs, and the same article with different headlines is still treated as one unique article.

⁴Some popular articles do not have text because they are in multimedia formats such as videos or questionnaires.

⁵All the statistics in this section count each article as many times as it appears in the daily data, i.e., they are the summary statistics of the observations, not unique articles in the M-Emailed and M-Facebook lists. However, these summary statistics are largely the same if we instead count each article only once per list.

followed by U.S (at 10%) and then Health, Well, Business Day, etc. In contrast, in the M-Facebook list, over 30% of the articles come from the U.S, followed by Opinion, World, and New York (see Figure A2 in Appendix §A for details). In summary, the length and section headings of articles that appear on the two lists are quite different.

Nevertheless, metadata such as section names and type of material have limited ability to categorize news content or explain the difference in the news content across the two lists for two reasons. First, newsworthy topics change regularly, and it is hard for rigid and long-established news structures (e.g., section names) to capture constantly evolving topics covered in news-cycles. Second, most news articles cover two or more topics, which makes it difficult to categorize them under one section name or news desk. Please see Appendix §A.2 for a detailed example with two articles that highlight these challenges. Therefore, we need to go beyond the meta-data and learn more about the content and tone of the articles to quantify the differences between the articles shared through the two different mediums.

3 Empirical Analysis and Results

Our empirical analysis consists of two steps. In the first step, we quantify the content of the articles using topic models and measures of polarization in §3.1 and §3.2. Next, in the second step in §3.3, we use the polarization scores and topics to model the relationship between how polarized an article is vs. how it is shared on Facebook vs. email.

3.1 Topic Modeling using LDA

Topic models help researchers organize and provide insights into large collections of unstructured text data. Latent Dirichlet Allocation (LDA) is the most common topic model, and it models each document (text) as a distribution over topics and each topic as a distribution over words (Blei et al., 2003). This allows documents to be a part of different topics, rather than being separated into discrete groups. LDA models have been extensively used in marketing; typically, researchers apply the LDA model (or derivatives of it) to derive topics and then use them as inputs in downstream models. The applications range from product reviews (Tirunillai and Tellis, 2014), social media content Zhong and Schweidel (2020), restaurant menus (Puranam et al., 2017), online search queries (Liu and Toubia, 2018), entertainment products (Toubia et al., 2019).

Table 1: Topics

Nº	Topic	Prevalence	Top 10 words
1	Family	0.067	family, home, friend, feel, child, old, mother, love, never, live
2	Politics	0.066	political, america, article, editor, hear, commit, letter, email, power, diversity
3	Emotions and Feelings	0.059	really, feel, lot, mean, kind, talk, start, happen, ask, question
4	Coronavirus Pandemic	0.049	virus, coronavirus, health, test, case, pandemic, covid, death, spread, infection
5	Books	0.043	book, write, story, world, read, writer, man, death, author, novel

Continued on next page

Table 1: Topics

Nº	Topic	Prevalence	Top 10 words
6	Architecture	0.042	open, place, street, room, city, old, design, house, th, century
7	Money, Personal Finance	0.040	pay, money, percent, tax, economic, job, worker, government, economy, income
8	New York City	0.031	city, york, county, home, resident, local, community, area, restaurant, ms
9	Music/Movies	0.031	music, play, film, movie, song, star, watch, character, series, theater
10	Health Research, Lifestyle Advice	0.030	study, dr, researcher, research, percent, university, scientist, risk, body, health
11	Nature	0.030	water, tree, fire, mile, island, river, area, park, foot, town
12	Black Lives Matter	0.029	police, officer, protest, protester, kill, man, death, video, arrest, fire
13	Women’s Issues, Sexual Harassment	0.029	ms, woman, interview, family, man, girl, sexual, member, sex, write
14	Donald Trump	0.029	trump, president, house, white, administration, news, washington, fox, donald, former
15	Elections	0.028	republican, election, vote, party, senator, house, democrat, senate, president, trump
16	Joe Biden	0.028	biden, campaign, democratic, candidate, voter, party, president, political, trump, presidential
17	Political Investigations	0.027	case, lawyer, investigation, charge, prosecutor, attorney, justice, report, office, department
18	Public Health and Medicine	0.025	patient, hospital, doctor, medical, health, care, drug, dr, treatment, die
19	Racial Identity and History	0.024	black, white, racial, african, race, man, history, woman, community, america
20	Business	0.024	company, business, executive, employee, industry, market, sell, product, chief, amazon
21	Social media	0.024	facebook, video, post, medium, online, app, social, twitter, datum, digital
22	Education, School System	0.023	school, student, child, parent, college, university, class, teacher, education, family
23	Supreme Court and Judicial System	0.022	court, law, justice, rule, judge, case, federal, supreme, legal, administration
24	Food	0.021	food, wine, restaurant, eat, cook, recipe, meat, add, dish, flavor
25	World News	0.017	european, world, europe, germany, britain, british, france, german, united, french
26	American Military	0.017	military, war, iran, force, united, official, general, troop, iraq, iranian
27	Russia	0.017	official, russia, russian, intelligence, security, ukraine, report, department, government, agency
28	Covid Vaccine	0.015	vaccine, dose, johnson, vaccination, health, agency, receive, federal, administration, government
29	China, India, International Travel	0.014	china, chinese, government, travel, united, india, flight, passenger, airport, airline
30	Real estate	0.014	home, building, estate, house, property, apartment, real, rent, buy, housing

Continued on next page

Table 1: Topics

Nº	Topic	Prevalence	Top 10 words
31	Power, Energy Supply, and Climate	0.014	climate, change, power, energy, oil, car, environmental, plant, gas, water
32	Sport	0.012	game, team, player, play, league, sport, season, coach, club, baseball
33	Art, planes	0.011	art, museum, artist, bird, plant, painting, plane, paint, pilot, boeing
34	Science	0.011	dr, science, space, scientist, university, human, lab, team, paper, earth
35	Covid protection	0.011	mask, wear, risk, face, hand, air, bike, safe, coronavirus, indoor
36	Israel	0.008	israel, gun, israeli, palestinian, jewish, muslim, group, jew, violence, attack
37	Christianity and Church	0.007	church, abortion, religious, christian, woman, catholic, gay, faith, god, evangelical
38	Horse Racing and Farms	0.004	farmer, farm, run, horse, japan, race, japanese, sport, runner, olympic
39	Pets and Animals	0.003	animal, dog, human, cat, specie, pet, wild, wildlife, park, fish
40	Judaism	0.002	jewish, funeral, smell, kelly, community, allen, wedding, brooklyn, rabbi, jew

We performed LDA on the corpus of 13508 articles using the Gensim package for Python, which is based on the variational Bayes algorithm described by Hoffman et al. (2010). Details of the data processing and hyperparameter tuning to identify the optimal number of topics are in Appendix B. Table 1 shows the summary of the 40 topics recovered by our LDA model, in decreasing order of prevalence in the corpus.⁶ In our data, the most prevalent topic is *Family* and the least prevalent is *Judaism*. The third column shows the top 10 words for each topic, listed in decreasing order of their share in that topic. The topic names were chosen manually by heuristically combining these keywords into a single phrase. For instance, Topic 4 contains such words as virus, coronavirus, health, test, case, and pandemic, and therefore was named *Coronavirus Pandemic*.

Overall, we find that the LDA model is able to uncover the latent topics in the corpus quite effectively. Figure 2 shows how a topic’s prevalence changes over time, for the ten most prevalent topics in the corpus. Notice that the *Pandemic* topic was almost non-existent till the end of 2019, but became the most popular topic at the beginning of 2020. We refer interested readers to Appendix §B.3 for additional results on the LDA analysis, including detailed word clouds of the pre-dominant words in each topic and links to the top three articles with the highest proportion of the topic (in our corpus).

3.2 Polarization Measures

Recall that our main research question asks whether the polarization of an article predicts the relative likelihood of being shared on Facebook vs. email. As such, a key concept that we need to define and measure is the political polarization of news content. Formally, a news article is considered politically polarizing if the content, text, and opinions expressed diverge away from the center and are closer to either of the extreme ends of the ideological spectrum (DiMaggio et al., 1996; Baldassarri and Gelman, 2008).

⁶Formally prevalence of topic j is defined as $prevalence_j = \frac{\sum_d p_{jd} \cdot \text{length}_d}{\sum_j [\sum_d p_{jd} \cdot \text{length}_d]}$, where d denotes a document and p_{jd} denotes the proportion of topic j in document d . Prevalence for all topics in a document sums to 1.

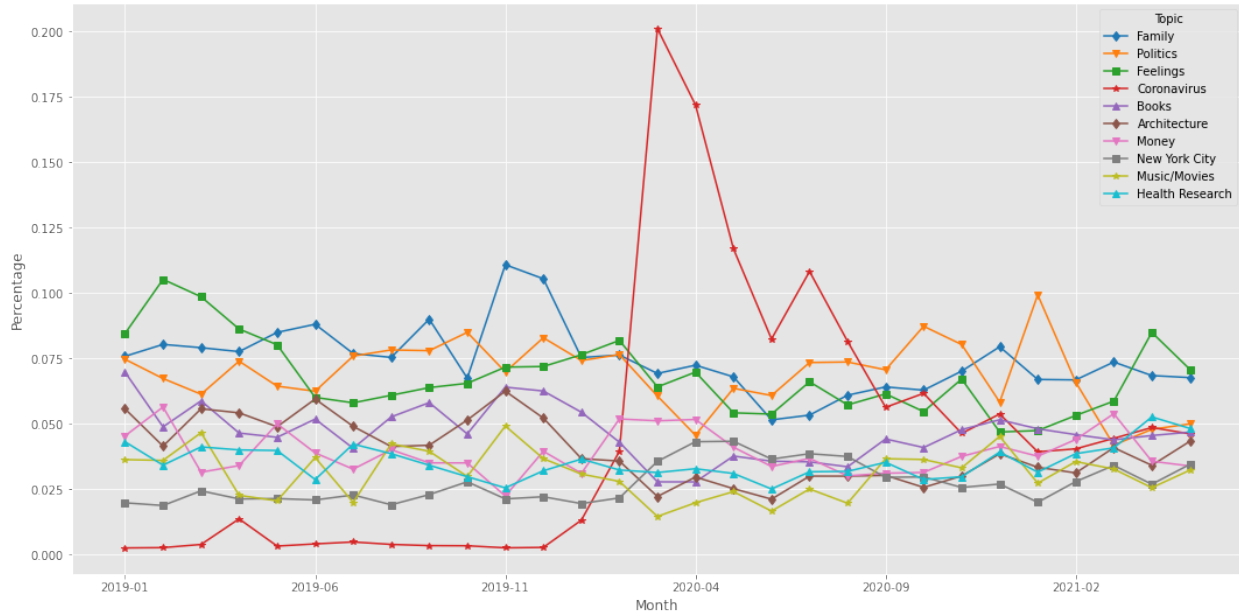


Figure 2: Topic prevalence over time

Prior research has used a variety of approaches to derive or predict the polarization of text/speech by an agent. Typically these approaches fall into two broad categories. In the first set of methods, the researcher has labeled data on the political party/affiliation of the agent or the outlet who created the text. Then, taking the affiliation of the agent/outlet as the ground truth, they characterize the differences in text or speech of the two parties/groups and use these differences to quantify the extent of polarization in a given piece of text; see (Gentzkow and Shapiro, 2010; Gentzkow et al., 2019). In the second set of methods, the researchers use crowd-sourced methods (Amazon MTurk) and surveys to score the slant/polarization of individual news articles (Budak et al., 2016). The former approach can only work when there are clear and well-known political affiliations for each piece of text, e.g., in the case of congressional speech, while the latter approach is not scalable beyond a small set of articles.

In this paper, we adopt a novel approach to measure political polarization that overcomes the scalability and lack of affiliation problems – we turn to the newly developed Large Language Models (LLMs) to obtain polarization scores and also validate these scores using standard surveys from human raters, whose scores can be considered an objective and true measure of human opinions on the polarization of content but are not scalable when the number of articles/content is high. We describe both approaches below.

3.2.1 LLM Measures of Polarization

We used GPT-3.5-turbo, a large language model, to generate polarization scores, and obtained three types of polarization scores (on a range of 1 to 5):

- Article-level polarization score – Here we provide the text of each article and ask the model to provide a polarization score.

- Topic-level polarization score – Here we provide the topic names derived from the LDA model and ask the model to provide a polarization score for each of the 40 topics.
- Topic keywords-level polarization score – One concern with the topic-level score is that the LLM may be very sensitive to the particular way we named the topics. Therefore, for robustness, we provided the model with only the top 10 keywords for each topic (and no topic name) as in Table 1), and obtained a new set of polarization scores.

We refer readers to Web Appendix §C for a detailed discussion of the prompts, the temperatures for each prompt (that drive the stochasticity of the answers), the number of iterations per question, the procedure used to standardize the polarization scores.

3.2.2 Survey Measures of Polarization

Next, we conducted a survey to measure the extent to which each of the topics identified from the LDA analysis are considered to be politically polarizing. Note that we use topic-level polarization measures here (instead of article-level) because the number of articles (and the length of each article) made it prohibitively expensive to obtain article-level scores. The subjects were undergraduate students at a large state university on the West Coast. Details of the demographics of the respondents and their news-reading habits are shown in Appendix §D.2. Survey respondents were presented with 10 random topics and were asked to rate how politically polarized the news coverage on a topic is, on a scale from 1 (Not at all polarized) to 5 (Extremely polarized) followed by questions about demographics and news reading and sharing habits. To understand people’s motivation for sharing news articles, we also asked respondents to rate how important it was for them that their social circle knew of their opinions on each of the 10 random topics presented to them. This tells us the extent to which sharing opinions on a topic is relevant from an identity-signaling perspective.

Table 2: Polarization measures (standardized)

	Survey	LLM_Topics_Avg	LLM_Keywords_Avg	LLM_Article
Family	-0.32	-1.23	-1.44	-0.97
Books	-0.82	-1.44	-1.33	-0.74
Feelings	-0.34	-0.69	-0.78	-0.61
Horse Racing	-0.93	-0.61	-0.98	-0.79
Music/Movies	-0.62	-1.36	-1.25	-1.32
Joe Biden	0.78	0.90	1.26	0.89
Money	-0.18	-0.69	0.53	0.44
Elections	0.79	1.28	1.26	1.67
Public Health	0.33	0.06	0.46	-0.66
Pets and Animals	-1.27	-1.44	-0.96	-0.98
Donald Trump	0.88	1.58	1.46	1.57
New York City	-0.33	-0.68	-0.98	-0.22
Women’s Issues	0.69	0.89	0.51	0.11
Architecture	-1.26	-1.37	-1.44	-1.43

Continued on next page

Table 2: Polarization Measures (standardized)

	Survey	LLM_Topics_Avg	LLM_Keywords_Avg	LLM_Article
Coronavirus	0.45	1.11	1.08	-0.07
Science	-0.12	-0.83	-0.69	-1.20
American Military	0.46	0.14	0.70	1.01
Food	-1.00	-1.28	-1.44	-1.94
Health Research	-0.46	-0.37	-0.03	-0.76
Business	0.20	-0.45	0.17	0.13
Social Media	0.08	0.07	0.05	-0.09
Black Lives Matter	0.72	1.26	0.98	1.06
Art, Planes	-0.95	-1.23	-1.23	-1.19
Racial Identity	0.57	1.12	1.07	0.74
World News	0.34	0.14	-0.60	0.14
Nature	-0.45	-1.44	-1.35	-1.27
Judicial System	0.57	0.67	0.44	1.40
Covid Vaccine	0.44	0.97	0.79	-0.07
China, India	0.25	0.42	0.62	0.26
Real Estate	-0.60	-0.69	-0.78	-0.54
Russia	0.14	0.67	0.81	1.58
Power and Climate	0.38	0.83	0.98	0.36
Politics	0.89	1.58	1.08	1.57
Sports	-0.90	-0.69	-0.96	-0.48
Political Investigations	0.77	0.81	0.71	1.32
Israel	0.33	1.04	1.07	1.59
Church	0.25	0.37	0.90	0.91
Covid Protection	0.36	0.53	0.06	-0.61
Education	-0.04	0.14	-0.39	-0.20
Judaism	-0.18	-0.08	-0.33	-0.58

3.2.3 Polarization Scores Summary

In Table 2, we present a summary of all four polarization scores. To aggregate the polarization scores from the article to the topic level (last column), we use the weighted average of the polarization score across all articles, where the weights are the proportion of the topic in a given article. We then correlate all the polarization scores and find that there is an extremely high correlation between survey measures of polarization and LLM-generated polarization scores. This is a useful finding because it suggests that future researchers can use LLMs to score news content and text on polarization and ideological issues. In our specific context, this finding allows us to use the LLM-generated article-level polarization scores in our empirical analysis. Further, Table A4 in Web Appendix §C.1 provides examples of the rationale that the LLM provides for its polarization scores. As we can see from this table, the LLM is quite good at explaining why it scores certain articles higher and others lower.

In terms of substantive findings, we see that topics such as *Politics, Elections, Joe Biden, Political*

Investigations, Black Lives Matter, Women’s Issues and Sexual Harassment, and Racial Identity and History are considered to be the most polarizing. On the other hand, topics such as *Pets and Animals, Architecture, Food, Horse Racing and Farms, Sports, and Books*. In §3.3, we examine this issue further and estimate the impact of polarization, *after* controlling for the topics in the article.

Table 3: Correlation Among the Different Polarization Scores

	Survey	LLM_Topics_Avg	LLM_Keywords_Avg	LLM_Article
Survey	1.000000	0.908675	0.880517	0.806447
LLM_Topics_Avg		1.000000	0.925010	0.836079
LLM_Keywords_Avg			1.000000	0.858370
LLM_Article				1.000000

3.3 Polarization and Sharing Behavior

3.3.1 Empirical model

We now specify a simple descriptive model to quantify the difference in the polarization of articles across the two lists. Recall that the rank is the position of an article in the M-Emailed or M-Facebook list, and can go from 1 (most popular) to 20 (least popular). We define Y_{it} as a measure of the difference between the rank of an article i in the M-Emailed list and the rank of an article i in the M-Facebook list on the day t . We consider two measures for Y_{it} : (1) a simple difference metric⁷, and (2) an indicator for whether the article was ranked higher in the M-Emailed list compared to the M-Facebook list. Therefore, lower values of Y_{it} indicate that article i is more popular in the M-Emailed list compared to the M-Facebook list.

Next, we specify Y_{it} as a function of the polarization score of article i and other controls as follows:

$$Y_{it} = \alpha + \beta P_i + \sum_{j=1}^{n-1} \gamma_j \cdot p_{ij} + \delta Article_i + \zeta Age_{it} + \epsilon_{it}, \quad (1)$$

where P_i is the polarization score of article i , p_{ij} is the proportion of topic j in article i . The proportion of all the topics in an article adds up to 1, so we exclude the proportion of the topic *Social media* to avoid collinearity. Since the total number of topics n is 40, we have $n - 1 = 39$ topics in the model. Next, $Article_i$ consists of article-specific attributes such as its length, the length of its headline, and the section name. We also include controls for the age of the article (defined as the number of days since release) because descriptive evidence suggests that there is a difference in the stickiness of articles across the two lists (see Figure A3 and the accompanying discussion in Web Appendix §A). Finally, for data, we use observations at the day-article level, where for each day t we include all articles that were ranked at least in one of the lists on that day. This gives us 20–40 observations for each day, which amounts to a total of 23580 observations over a period of 697 days.

⁷If article i is not ranked in the top-20 in a given list on day t , then we specify its rank as 25 for that list to calculate y_{it} . In Appendix E, we present robustness checks with other numbers.

Since Y_{it} is the difference in the popularity of article i across the two lists on day t , it differences out common time-specific shocks that affect an article’s popularity. For example, if the topic *Joe Biden* was popular during elections, then articles on this topic will appear in both M-Emailed and M-Facebook lists. Thus, Y_{it} captures the incremental popularity of the article on email (compared to Facebook) after controlling for other time-varying shocks to the article’s popularity. Further, this specification captures the impact of polarization after controlling for the topic distribution of the article. Thus, differences in topics’ inherent tendency to be shared via social media vs. shared privately, are already captured/controlled for. Further, even if some topics are more polarizing than others (as shown in Table 1b), this captures the effect of within-topic variation in polarization on Y_{it} .

3.3.2 Results and Discussion

Table 4 shows the regression results, where Model (1) uses the difference in ranks as the dependent variable and Model (2) uses the binary indicator as the outcome variable. In both regressions, we see that the polarization score has a positive coefficient, which means that more polarizing articles are more commonly shared on Facebook, compared to email. Note that since we control for the prevalence distribution of the topics in each article, this estimate is the incremental impact of polarization after controlling for the topic distribution of the article (and the effect of the topic on the tendency to be shared on the two mediums). In Web Appendix §E, we consider a model where we do not control for topics, and find that the effect of polarization is overestimated in that case. This is understandable since topics that tend to have more polarized reporting also tend to be shared more on Facebook; and also emphasizes the importance of controlling for topics. Further, in Web Appendix §E, we present a series of robustness checks to show that these results are valid even when we vary the model specification and control for the emotional content of the article using LIWC measures (Boyd et al., 2022; Berger and Milkman, 2012).⁸

In addition, we find that certain topics are more likely to be posted on social media compared to being shared via email (after controlling for the polarization score and other control variables). Specifically, articles on topics such as *Books, Business, Animals, Food, Real-estate, Nature, and Health Research and Lifestyle Advice* are more commonly sent through email. In contrast, articles on topics such as *Election investigations, Covid vaccine, Russia, Women’s Issues and Sexual Harassment, Coronavirus Pandemic, Black Lives Matter* etc., are more likely to be posted on social media. While we do not take a stance on why certain topics are shared more widely on social media, it is worthwhile to note that these patterns are consistent with some natural explanations and earlier works. Notice that the topics shared on social media tend to be of broader interest (e.g., politics, elections), and, understandably, they are shared on social media with a larger set of acquaintances on Facebook compared to other topics. Further, we find that the correlation between a topic’s social signaling score (see Table A5 in Appendix §D.2) and its coefficient from the regression results of Model (1) is 0.35. This positive correlation is consistent with earlier work that suggests that identity signaling

⁸Berger and Milkman (2012) show that content that evokes high-arousal positive (awe) or negative (anger or anxiety) emotions is more likely to be shared by email, whereas content that evokes low-arousal, or deactivating, emotions (e.g., sadness) is less likely to be shared. Like us, they also use data from the New York Times. However, they confine their analysis to the most e-mailed articles, while our goal is to contrast the differences in the sharing patterns of news articles across social media and emails.

regarding can be a strong motivator behind users' actions on social media (Berger, 2008; Reed II et al., 2012; Reed II and Forehand, 2019; van der Does et al., 2022).

Table 4: Results from Equation 1 capturing the difference in the polarization of articles across the two lists.

Dependent Variables: Model:	Difference in ranks (E-F)		Higher in F (binary)	
	(1) OLS		(2) OLS	
<i>Variables</i>				
Polarization score	1.088***	(0.108)	0.041***	(0.004)
Family	12.783***	(1.639)	0.467***	(0.056)
Politics	-4.377**	(1.853)	-0.184***	(0.062)
Emotions and Feelings	5.542***	(1.809)	0.204***	(0.059)
Coronavirus Pandemic	18.352***	(1.554)	0.629***	(0.055)
Books	-2.659	(1.983)	-0.098	(0.068)
Nature	2.017	(1.843)	0.029	(0.061)
Women's Issues, Sexual Harassment	27.063***	(1.954)	0.915***	(0.069)
Business	-4.341*	(2.399)	-0.195**	(0.082)
Education, School System	3.072	(1.895)	0.080	(0.066)
American Military	15.975***	(1.780)	0.520***	(0.061)
China, India, International Travel	23.807***	(2.263)	0.793***	(0.080)
Power, Energy Supply, and Climate	6.178***	(2.141)	0.187**	(0.076)
Judaism	-26.599***	(6.431)	-1.226***	(0.250)
Architecture	-12.896***	(1.849)	-0.545***	(0.063)
Money, Personal Finance	7.010***	(1.598)	0.228***	(0.054)
New York City	7.669***	(2.047)	0.266***	(0.072)
Music/Movies	10.841***	(2.089)	0.393***	(0.074)
Health Research, Lifestyle Advice	-3.980**	(1.768)	-0.132**	(0.058)
Black Lives Matter	21.575***	(1.720)	0.682***	(0.055)
Donald Trump	14.857***	(1.876)	0.417***	(0.065)
Elections	13.087***	(1.688)	0.425***	(0.056)
Joe Biden	11.072***	(1.750)	0.359***	(0.061)
Political Investigations	15.250***	(1.571)	0.459***	(0.054)
Public Health and Medicine	7.452***	(1.822)	0.255***	(0.063)
Racial Identity and History	6.107***	(1.903)	0.173**	(0.068)
Supreme Court and Judicial System	11.040***	(1.852)	0.381***	(0.065)
Food	-2.080	(1.813)	-0.080	(0.062)
Covid Vaccine	17.559***	(1.799)	0.508***	(0.063)
Art, Planes	-5.755**	(2.536)	-0.216**	(0.089)
Covid Protection	-3.213*	(1.899)	-0.181***	(0.068)
Christianity and Church	6.937**	(2.730)	0.198*	(0.103)
Horse Racing and Farms	6.992	(4.390)	0.267*	(0.155)
World News	2.044	(2.351)	0.026	(0.084)
Russia	9.526***	(1.967)	0.322***	(0.068)
Real Estate	-7.314***	(2.318)	-0.244***	(0.081)

Continued on next page

Table 4: Results from Equation 1

Dependent Variables: Model:	Difference in ranks (E-F)		Higher in F (binary)	
	(1) OLS		(2) OLS	
Sports	8.431***	(2.669)	0.188**	(0.087)
Science	4.450	(2.723)	0.082	(0.097)
Pets and Animals	-21.780***	(4.689)	-0.758***	(0.168)
Israel	5.820*	(2.966)	0.198**	(0.100)
Headline length	0.087***	(0.033)	0.003***	(0.001)
Snippet length (standardized)	0.558***	(0.095)	0.019***	(0.003)
Word Count (standardized)	-1.152***	(0.094)	-0.041***	(0.003)
In Print (binary)	-1.863***	(0.194)	-0.071***	(0.007)
(Intercept)	-22.660***	(1.566)	-0.243***	(0.053)
<i>Controls</i>				
Days after release (quadratic)		Yes		Yes
<i>Fixed-effects</i>				
Section name		Yes		Yes
<i>Fit statistics</i>				
Observations		23,580		23,580
R ²		0.33600		0.32589
Adjusted R ²		0.33346		0.32331

Clustered (date_id) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

We now provide some additional discussion and interpretation of the results. First, our results are descriptive and should be interpreted carefully. Since we do not manipulate polarization or slant within an article exogenously, our results do not state that certain polarization causes an article to be more/less likely to be posted on Facebook. Rather it simply states, that after controlling for a series of observables such as the topics covered, length, news section, and emotional content of an article, polarizing articles are more likely to be shared through Facebook rather than email. That said, the observed patterns may stem from differences in the segments of consumers who post on social media vs. those who share on email or differences in how the same users employ the two communication media (the same user may share cerebral articles with close friends through email but post political/polarizing content on Facebook). Alternatively, these patterns may also reflect the implicit effect of social media algorithms. That is, Facebook users may have learned that their posts on more polarizing topics are more likely to be popular and/or amplified by the internal algorithm and hence favor those types of articles when posting on Facebook.

Nevertheless, we believe that documenting these descriptive results can help further discussion on this topic, and future research could further examine the sources and channels of polarized content.

3.3.3 Did Sharing Behavior Change over Time?

Table 5: Average polarization scores (standardized) of articles that appeared at least once in the Most emailed and Most shared on Facebook lists in the pre- and post-election periods.

Period	Sample	Avg. Polarization
Pre	Emailed	-0.08
Post	Emailed	-0.22
Pre	Facebook	0.21
Post	Facebook	-0.00

Recent research has shown that polarization in preferences, behavioral intentions, and actual purchase decisions for consumer brands increased after the election of Donald Trump in 2016 (Schoenmueller et al., 2022). Moreover, polls and anecdotal evidence suggest that polarization on social media platforms has exacerbated after the 2020 elections (Jurkowitz et al., 2020). Indeed, discussions about the polarization of social media platforms have gained urgency and prominence after the Jan. 6 assault on the U.S. Capitol. Experts have argued that these incidents were fomented by the divisive discourse on social media and these issues have been the subject of a recent senate investigation (Reuters-Staff, 2021).

Motivated by these arguments, we examine whether the sharing patterns are different after the 2020 elections. We consider two subsets of our data – (1) Pre – data from January 1, 2020, to October 30, 2020, and (2) Post – data from December 1, 2020, to May 30, 2021. We then run the same model (as shown in Equation (1), with the difference in ranks as the outcome variable) but also include a *Post* variable and the interaction between *Post* and polarization. We find that: (1) the main effect of polarization continues to be positive, i.e., more polarized articles have a higher relative likelihood of being shared on Facebook in both periods and (2) the interaction effect is negative, i.e., the polarization score of articles plays a smaller role in predicting the differences across the two lists after the election (see Appendix §F for details). To further understand these patterns, we summarize the average polarization scores of articles that appeared at least once in the Most emailed and Most shared on Facebook lists in the pre- and post-election periods in Table 5. Interestingly, we see that the articles shared on both channels were less polarized after the elections, though this drop is higher for Facebook (which explains the results from the regression). In summary, at least in this setting, we do not find there was any significant increase in polarization of articles shared on social media after the elections (compared to email).

4 Conclusion

In this paper, we examine if and how the content of articles seeded on social media (specifically, Facebook) differs from those sent via email. We use data from the New York Times Most Emailed and Most Shared on Facebook lists for a 2.5-year period for our study. For each article, we recover the topic distribution using LDA and the polarization score using LLMs, and connect the difference in the article’s ranking across the two lists with its polarization score and topic. We show that more polarizing articles are more likely to be seeded on social media (compared to email), after controlling for a series of confounding factors such as the topic, news section, emotion, age, etc. Our results are descriptive and should be interpreted as summarizing

sharing patterns on different channels, and not as the causal effects of polarization on users' sharing behavior.

Our analysis comes with a set of caveats, which can serve as avenues for further research. First, since we do not observe individual-level data, we cannot comment on whether the same user shares different content across the two media formats or whether the set of users posting on Facebook is systematically different from those who share news through email. Second, while our analysis shows that these patterns exist before the explicit impact of Facebook's algorithms, it is not clear if there is an implicit impact. It would be useful to examine whether users post more polarizing articles on Facebook anticipating that such articles will be more popular (because Facebook's algorithm promotes such articles), or if this behavior is purely exogenous. Studies that separate the explicit and implicit role of algorithms on user behavior would be an excellent next step. Third, the findings are specific to the setting that we study, i.e., NYTimes readers who share articles on Facebook and/or email. Future research that expands that scope to other news websites and social media websites can help with establishing the generalizability of these patterns. Finally, while we do not delve too much into the incentives of news platforms, future research could build on our findings and the growing analytical work that examines platform and news aggregator incentives to create content and price their products (Amaldoss et al., 2021; Amaldoss and Du, 2023).

References

- H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–76, March 2020. doi: 10.1257/aer.20190658. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20190658>.
- W. Amaldoss and J. Du. Express: How can publishers collaborate and compete with news aggregators? *Journal of Marketing Research*, page 00222437231153607, 2023.
- W. Amaldoss, J. Du, and W. Shin. Media platforms' content provision strategies and sources of profits. *Marketing Science*, 40(3):527–547, 2021.
- C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
- E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015. doi: 10.1126/science.aaa1160. URL <https://www.science.org/doi/abs/10.1126/science.aaa1160>.
- D. Baldassarri and A. Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.
- P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.
- J. Berger. Identity signaling, social influence, and social contagion. *Understanding peer influence in children and adolescents*, pages 181–199, 2008.
- J. Berger and K. L. Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2): 192–205, 2012. doi: 10.1509/jmr.10.0353. URL <https://doi.org/10.1509/jmr.10.0353>.

- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning research*, 3 (Jan):993–1022, 2003.
- L. Boxell, M. Gentzkow, and J. M. Shapiro. Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617, 2017.
- R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 2022.
- J. Brand, A. Israeli, and D. Ngwe. Using gpt for market research. *Available at SSRN 4395751*, 2023.
- C. Budak, S. Goel, and J. M. Rao. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly*, 80(S1):250–271, 04 2016. ISSN 0033-362X. doi: 10.1093/poq/nfw007. URL <https://doi.org/10.1093/poq/nfw007>.
- M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.
- N. Dev Portal. Article search api. *NYTimes*, 2022. URL <https://developer.nytimes.com/docs/articlesearch-product/1/overview>.
- P. DiMaggio, J. Evans, and B. Bryson. Have american’s social attitudes become more polarized? *American journal of Sociology*, 102(3):690–755, 1996.
- G. Eady, J. Nagler, A. Guess, J. Zilinsky, and J. A. Tucker. How many people live in political bubbles on social media? evidence from linked survey and twitter data. *Sage Open*, 9(1):2158244019832705, 2019.
- M. Gentzkow and J. M. Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.
- M. Gentzkow and J. M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.
- M. Gentzkow, J. M. Shapiro, and M. Taddy. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340, 2019. doi: <https://doi.org/10.3982/ECTA16566>. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16566>.
- A. Goli and A. Singh. Language, time preferences, and consumer behavior: Evidence from large language models. *arXiv preprint arXiv:2305.02531*, 2023.
- M. Hoffman, F. Bach, and D. Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.
- M. Jurkowitz, A. Mitchell, E. Shearer, and M. Walker. U.s. media polarization and the 2020 election: A nation divided. *Pew Research Center*, 2020. URL <https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/>.
- R. Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70, March 2021. doi: 10.1257/aer.20191777. URL <https://www>.

aeaweb.org/articles?id=10.1257/aer.20191777.

- J. Liu and O. Toubia. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):930–952, 2018.
- E. Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.
- N. Persily and J. A. Tucker. Social media and democracy: The state of the field, prospects for reform. 2020.
- D. Puranam, V. Narayan, and V. Kadiyali. The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science*, 36(5):726–746, 2017.
- A. Reed II and M. Forehand. *Handbook of Research on Identity Theory in Marketing*. Edward Elgar Publishing, 2019.
- A. Reed II, M. R. Forehand, S. Puntoni, and L. Warlop. Identity-based consumer behavior. *International Journal of Research in Marketing*, 29(4):310–321, 2012.
- Reuters-Staff. Facebook does not believe it is a primary cause of polarization - exec to cnn. Reuters, 2021. URL <https://www.reuters.com/article/facebook-whistleblower-clegg-idCAKBN2GT0EO>.
- V. Schoenmueller, O. Netzer, and F. Stahl. Polarized america: from political partisanship to preference partisanship. *Frontiers, Marketing Science*, 2022.
- D. Shin and B. Kadiyala. Social learning with polarized preferences on content platforms. *Available at SSRN 3916284*, 2022.
- C. R. Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton University Press, 2018.
- S. Tirunillai and G. J. Tellis. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of marketing research*, 51(4):463–479, 2014.
- O. Toubia, G. Iyengar, R. Bunnell, and A. Lemaire. Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, 56(1):18–36, 2019.
- T. van der Does, M. Galesic, Z. O. Dunivin, and P. E. Smaldino. Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, 119(10):e2117898119, 2022.
- N. Zhong and D. A. Schweidel. Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science*, 39(4):827–846, 2020.

Web Appendix

A Appendix for Data Collection and Summary Statistics

A.1 Data Details and Summary Statistics

	Count	Unique values	Description
Link	13688	13688	Latest web url of the article
Main headline	13688	13684	Headline used on the article’s webpage
Abstract	13552	13527	Abstract of the article
Lead Paragraph	13457	12991	First paragraph of the article
Print section	10529	33	Section of the newspaper if article is printed: A, AR, AU, B, BR, BU, C, D, E, ED, F, L, LI, M2, MB, MC, MM, P, ...
Source	13554	4	Source of the article: The New York Times, International New York Times, International Herald Tribune, Wirecutter
Publication Date	13554	13139	Date of the publication
Document type	13554	3	Type of the document: article, multimedia, wirecutter article
Section name	13553	47	Section of the NYTimes: World, U.S., Politics, N.Y., Business, ...
Type of material	13553	16	Type of material: Op-Ed, News, Obituary (Obit), Interactive Feature, ...
Byline original	13258	4970	Author of the article
Full text	13508	13508	Full text of the article

Table A1: Description of Metadata fields

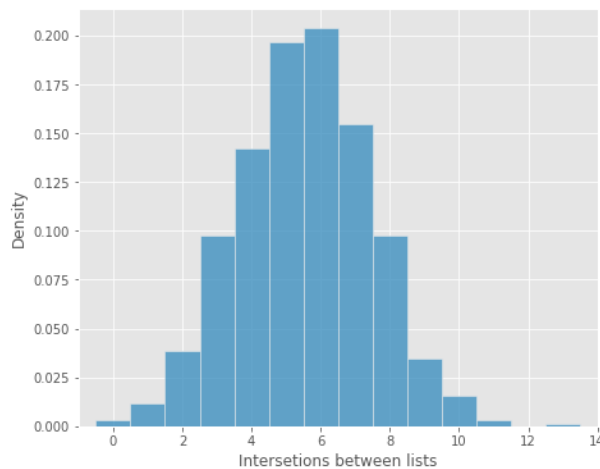
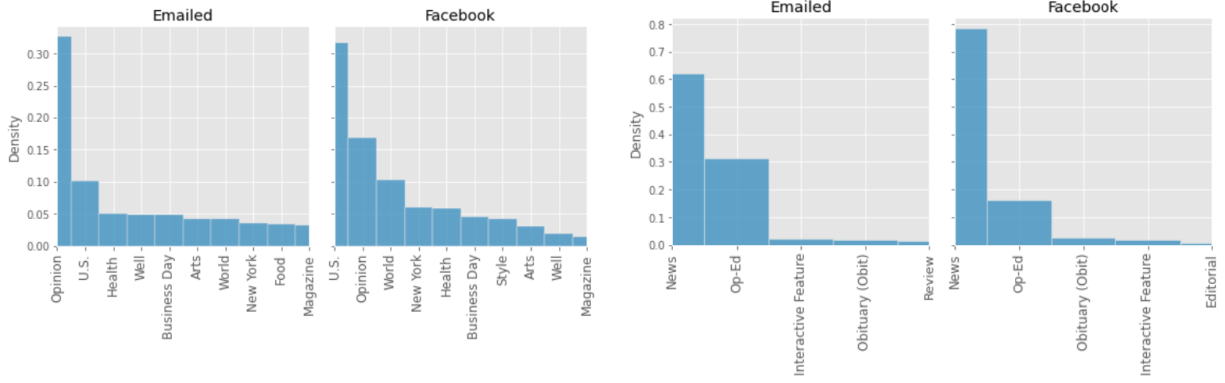


Figure A1: Daily overlap per day (number of identical articles) across the M-Emailed and M-Facebook lists.

In Figure A3, we present the distributions of the number of days an article spends in the top-20 list for both types of sharing. On average, articles spend 1.5 days in the M-Emailed list and 1.9 days in the M-Facebook list. Thus, there is more stickiness or persistence in the M-Facebook list compared to the M-Emailed list. The fact that the M-Facebook list is more sticky is corroborated in another way too: Over the entire observation period, the M-Emailed lists have a total of 9324 unique articles while the M-Facebook lists



(a) Section Name (b) Type of Material

Figure A2: Distributions of Section Names and Type of Material

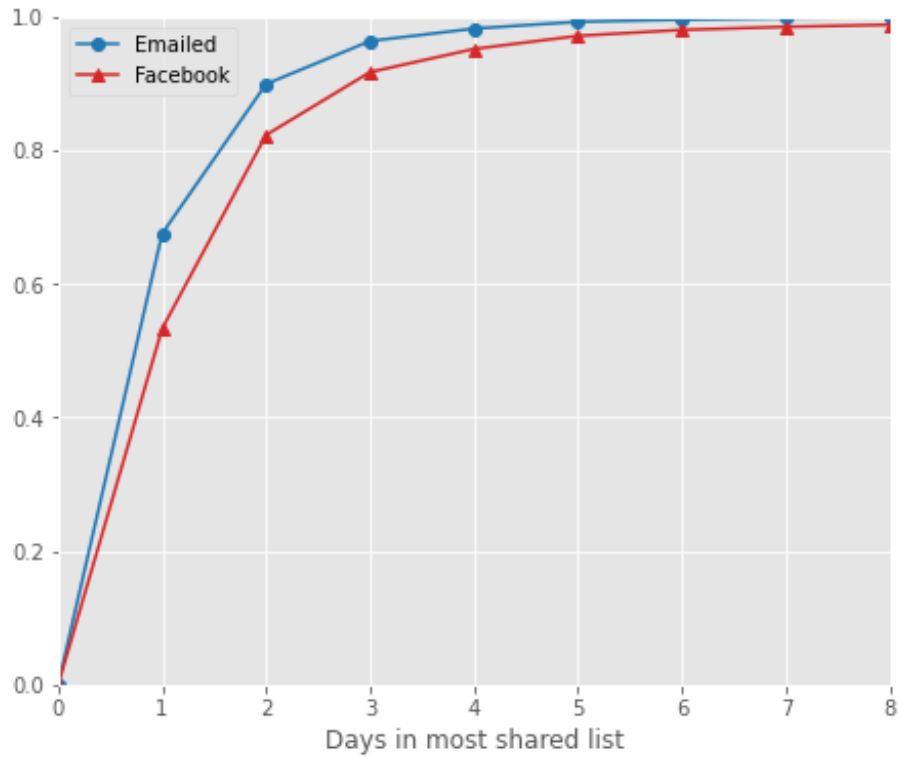


Figure A3: The stickiness of articles in the two lists, as captured by the CDF of the number of days an article stays on that list.

have only 7288 unique articles.

A.2 Example of Two Articles

Table A2 illustrates the challenges of relying only on metadata using an example with two articles. First, even though both articles are predominantly about Covid, the first article is placed under the *World* section and *National* news desk, while the second article is categorized under *Health* and *Science*. This is despite the fact that both articles were released many months after the outbreak, and news editors had sufficient time to learn how to categorize Covid-related articles. Second, notice that both articles cover multiple topics. For instance, the second article in Table A2 categorized under *Health* and *Science* also covers national policy, as can be inferred from the phrase “nationwide restrictions” in the lead paragraph. In addition, it is also focused on the United States and could have been easily categorized under *National*. Thus, this article can be represented as a mixture of different topics, such as a pandemic, health, policy, and national.

In contrast, our LDA model suggests that the *Coronavirus Pandemic* topic is the most prevalent topic in both articles, followed by the *Covid Vaccine* topic. Thus, the LDA model would classify both these articles as being very similar, unlike the metadata which classified them under different section headings.

Table A2: Example illustrating how two articles on the same topic (covid) can be categorized under different section names, and news desk.

Article 1	
Headline	New York reports its first case of the troubling variant that is dominant in Brazil.
Lead paragraph	New York has joined a growing list of more than a dozen states that have confirmed at least one case of a worrisome coronavirus variant first found in Brazil.
Section, News desk, Type of material	World, National, News
Article 2	
Headline	A Dose of Optimism, as the Pandemic Rages On
Lead paragraph	On March 16, back when White House news conferences were still deemed safe to attend, President Trump stood before reporters and announced that drastic nationwide restrictions — in schools, work places, our social lives — were needed to halt the coronavirus.
Section, News desk, Type of material	Health, Science, News

B Appendix for LDA Analysis

B.1 Data Pre-processing

Before estimating the model, we pre-process the data using some standard procedures for text analysis. First, we remove the set of high-frequency English words provided by Natural Language Tool Kit (NLTK) to generate a more meaningful vocabulary (Bird et al., 2009). Next, we create bigrams and trigrams to take into account combinations of words that often appear together. For instance, bigrams include the “United States”, “middle class”, or “prime minister”, while “Kim Jong-un” (hyphen excluded) is a trigram. Third, we use the Spacy library to lemmatize all words in the vocabulary. This process groups together the inflected forms of a word and replaces it with the base form, i.e., the lemma. For example, both words “modeled” and “modeling” will be replaced with the same lemma “model”. Finally, we remove all the words that appear in less than 15 articles or more than 50% of the corpus. In total, we obtain the vocabulary of 20668 words for the corpus of 13508 texts including 56 trigrams and 615 bigrams. Then, we train our model on the full set of 13508 articles.

B.2 Hyper-parameter tuning

To perform hyper-parameter tuning, we need a measure of model performance. In topic models, this measure is usually a metric that captures how interpretable the results from the model are (for human readers). These measures are usually referred to as “coherence scores” of the model, and many coherence scores have been proposed in the literature. For our analysis, we pick the coherence score recommended by Röder et al. (2015). This score has been shown to be the most correlated with human rankings in text analysis tasks, and is widely available in all topic model packages (including the Python Gensim package). The model has multiple hyper-parameters which require tuning, and the main one of interest is the number of topics. This hyper-parameter is critical since it has a big impact on the interpretability of the model results. Hence, we first focus on the hyper-parameter.

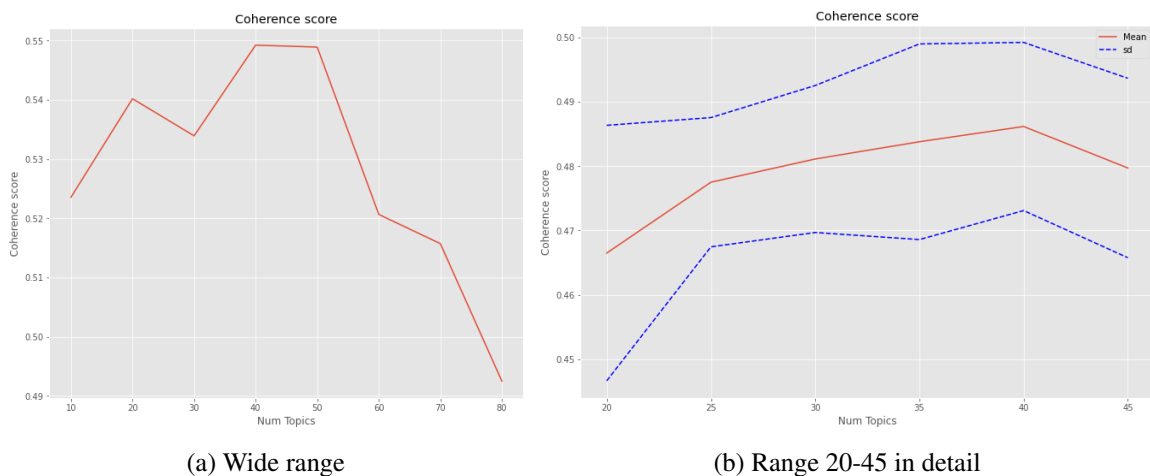


Figure A4: Coherence score for different number of topics

We start by exploring the model’s performance (as captured by the coherence score) as a function of the number of topics, while keeping all the hyper-parameters at their default values.⁹ In this exercise, we vary the number of topics from 10 to 80 (in increments of 10), and present the results in Figure A4a. As we can see,

⁹The only other adjusted parameters, apart from the number of topics, are a small number of passes through the corpus (10) for saving computational power and a smaller chunk size (100) for better performance with a small number of passes.

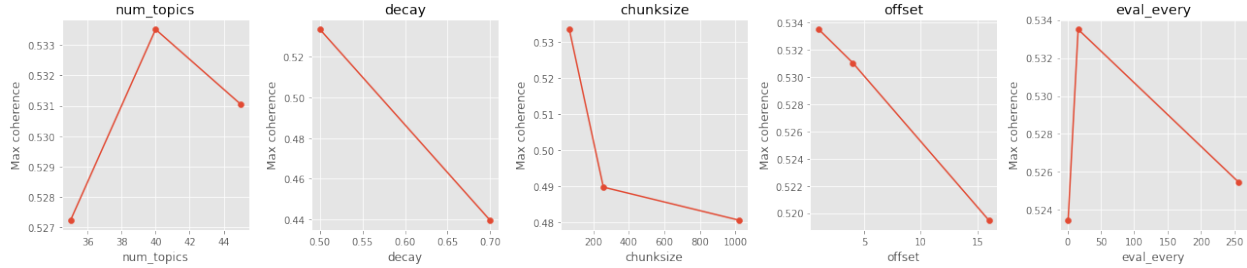


Figure A5: Max coherence score for all tuning parameters from a grid search

the performance of the model is low at 10 and also drops sharply after 50 topics. That is, if we include more than 50 topics or less than 20 topics, the recovered topics are not coherent and/or meaningful. Therefore, next exercise, we limit the range of topics from 20 to 45, but we perform a rigorous comparison in this range. Specifically, we consider smaller increments (of 5), and we run the model five times at each increment. This gives us measure of the average coherence of the model at a given topic number as well as some measure of variance in the model’s performance. The results from this exercise are shown in Figure A4b. We find that the range of 30-45 topics provides a better coherence score, although the variation of the estimates with the default setting is quite high. Thus, we now turn to tuning the other hyper-parameters in this range and jointly identify both the optimal number of topics as well as the other hyper-parameters.

Next, we perform a small grid search for the other hyper-parameters while a number of topics in the range of 30-45. Together, the set of hyper-parameters that need tuning are – (1) number of topics, (2) decay that weights how much information is forgotten when a new document is examined, (3) the size of a training chunk, (4) offset that controls the slowdown for the first steps, and (5) a parameter that states how often perplexity is evaluated. Based on a comprehensive grid search across all these hyper-parameters, best coherence score was ≈ 0.53 and was achieved with the number of topics equal to 40, decay of 0.5, chunksize 64, offset 1, and evaluated every 16 (see Figure A5). Note that this two-step process saves significant compute time compared to a large grid-search over all the hyper-parameters simultaneously.

There were other combinations of parameters that also resulted in almost the same coherence score. These combinations include the number of topics of 35 and 45, which suggests that the number of topics around that level does not have a significant effect on the coherence score. We chose 40 topics as it results in one of the best coherence scores, and it is sufficient to capture important topics. The final model was fitted with the parameters stated above and a high number of passes through the corpus (100). This helped us achieve a high coherence score of around 0.6 for the final model.

B.3 Additional LDA results



Figure A6: Wordclouds for LDA topics sorted by prevalence

Table A3: Examples: Articles with highest proportions of topics

N ^o	Topic	Most relevant article	Second relevant article	Third relevant article
1	Family	Tiny Love Stories: 'We Were Never Going to Venmo the Same \$15 Again'	Tiny Love Stories: 'It Was Unsettling to Find Myself Still Single at 36'	Tiny Love Stories: 'What Will You Miss if You Leave Him?'
2	Politics	The Two Crises of Conservatism	The Politics of Dystopia	No, Not Sanders, Not Ever
3	Feelings	An Adult's Guide to Social Skills, for Those Who Were Never Taught	The 4 'Attachment Styles,' and How They Sabotage Your Work-Life Balance	Transcript: Ezra Klein Interviews Judson Brewer
4	Pandemic	U.S. Surpasses 150,000 Coronavirus Deaths, Far Eclipsing Projections	C.D.C. Confirms First Possible Community Transmission of Coronavirus in U.S.	New York reports its first case of the troubling variant that is dominant in Brazil.
5	Books	Harold Bloom, Critic Who Championed Western Canon, Dies at 89	25 Great Writers and Thinkers Weigh In on Books That Matter	National Book Critics Circle Names 2020 Award Winners
6	Place	Make a Card, Any Card	36 Hours in Milan	In Southeastern Sicily, Old World Architecture Meets Stunning Beaches
7	Money	New Stimulus Package Brings Big Benefits to the Middle Class	90 million stimulus payments have gone out, the Treasury says.	How the Pandemic Has Changed Your Taxes
8	New York City	Driving and Travel Restrictions Across the United States	The Pizza-and-Beer Train: New York City's Hidden Railroad	N.Y. Will Move Homeless Men From Liberal Neighborhood After Backlash
9	Music/Movies	2021 Grammys Winners: The Full List	'Schitt's Creek' Sets an Emmy Record	Fran Drescher Developing 'The Nanny' Musical for Broadway
10	Research	Where You Carry Body Fat May Affect How Long You Live	The Best Time of Day to Exercise for Metabolic Health	A Single Session of Exercise Alters 9,815 Molecules in Our Blood
11	Nature	Wildfires Blot Out Sun in the Bay Area	Montana Hunkers Down as Snow Piles Up by the Foot. In September.	11 Great Alternatives to the Top National Parks
12	Black Lives Matter	Fires and Pepper Spray in Seattle as Police Protests Widen Across U.S.	A Capitol Police officer dies from injuries sustained during the pro-Trump rampage.	Police Dressed as Protesters: How Undercover Police in Hong Kong Severely Injured People
13	Women	Corroborating E. Jean Carroll	Why is Archie not a prince? Meghan says there's more to it than following the rules.	Sharon Osbourne Leaves 'The Talk,' CBS Says
14	Donald Trump	Kellyanne Conway, Long-Serving Trump Aide, Is Leaving the White House	Trump and his wife received coronavirus vaccine before leaving the White House.	Jen Psaki, Biden's new press secretary, pledges to bring 'truth and transparency back.'
15	Election	Trump Asked Pennsylvania House Speaker About Overturning His Loss	Impeachment Results: How Democrats and Republicans Voted	The Electoral College Is Voting Today. Here's What to Expect.
16	Joe Biden	Bill de Blasio Endorses Bernie Sanders	Pete Buttigieg Drops Out of Democratic Presidential Race	Bernie Sanders Wins Nevada Caucuses, Strengthening His Primary Lead

Continued on next page

Table A3: Examples

N ^o	Topic	Most relevant article	Second relevant article	Third relevant article
17	Investigation	Roger Stone and Everyone Charged in the 2016 Election Investigations	Paul Manafort Seemed Headed to Rickers. Then the Justice Department Intervened.	Michael Flynn Changes His Story, Putting Him on Collision Course With Judge
18	Health	The Infection That's Silently Killing Coronavirus Patients	The Risks of the Prescribing Cascade	Vaping Sicknesses Rising: 153 Cases Reported in 16 States
19	Racism	Black Like Kamala	Americans Are Divided by Their Views on Race, Not Race Itself	A Hard Conversation for the Latino Community
20	Business	Jeff Bezos to Step Down as Amazon C.E.O.	Prime Power: How Amazon Squeezes the Businesses Behind Its Store	The Great Amazon Flip-a-Thon
21	Social Media	Protecting Your Internet Accounts Keeps Getting Easier. Here's How to Do It.	How to Make the Most of Apple's New Privacy Tools in iOS 13	5 Tips to Take Command of Your iPhone
22	School	Colleges Plan to Reopen Campuses, but for Just Some Students at a Time	Online Schools Are Here to Stay, Even After the Pandemic	Interest Surges in Top Colleges, While Struggling Ones Scrape for Applicants
23	Law	Supreme Court Says Rejected Asylum Seekers Have No Right to Object in Court	The Supreme Court Is Making New Law in the Shadows	Trump Can't Immediately End DACA, Supreme Court Rules
24	Food	How to Make One-Pot Pasta With Practically Any Pasta	Make Cornmeal Muffins for Your Next Breakfast Treat	Our Best Vegan Thanksgiving Recipes
25	World	Boris Johnson Might Break Up the U.K. That's a Good Thing.	For Quebec, a French Woman May Not Be French Enough	Sweden Tries Out a New Status: Pariah State
26	Military	Pullback Leaves Green Berets Feeling 'Ashamed,' and Kurdish Allies Describing 'Betrayal'	President Endorses Turkish Military Operation in Syria, Shifting U.S. Policy	Fighting Next to U.S. Commandos, but Without the Same Training and Gear
27	Russia	Ukraine Knew of Aid Freeze by Early August, Undermining Trump Defense	Russian Intelligence Hackers Are Back, Microsoft Warns, Aiming at Officials of Both Parties	Russian Interference in 2020 Included Influencing Trump Associates, Report Says
28	Vaccine	Factory Mix-Up Ruins Up to 15 Million Vaccine Doses From Johnson & Johnson	Johnson & Johnson's vaccine is delayed by a U.S. factory mixup.	Trump administration officials passed when Pfizer offered months ago to sell the U.S. more vaccine doses.
29	China, India	British Airways Flight to Germany Lands in Scotland by Mistake	As Virus Spreads, U.S. Temporarily Bars Foreigners Who've Visited China	The Flight Goes Nowhere. And It's Sold Out.
30	Real Estate	It's a Buyers' Market for Two-Bedrooms	Home Sales Surge in Brooklyn	Real Estate Sales Continue to Stagnate in Manhattan
31	Climate	How Does Your State Make Electricity?	5 Global Trends Shaping Our Climate Future	As Coal Fades in the U.S., Natural Gas Becomes the Climate Battleground

Continued on next page

Table A3: Examples

N ^o	Topic	Most relevant article	Second relevant article	Third relevant article
32	Sport	Patrick Mahomes Flexes Chiefs' Muscle to Beat the Ravens	Tom Brady and the Patriots Are Upset by the Titans in a Stunner	Fielder's Choice? Cubs Turn a Simple Grounder Into a Wild Run.
33	Art, Planes	Confusion, Then Prayer, in Cockpit of Doomed Lion Air Jet	In Test of Boeing Jet, Pilots Had 40 Seconds to Fix Error	Boeing's 737 Max: 1960s Design, 1990s Computing Power and Paper Manuals
34	Science	Hunting for a Giant Black Hole, Astronomers Found a Nest of Darkness	Darkness Visible, Finally: Astronomers Capture First Ever Image of a Black Hole	Geminids Meteor Shower 2020: Watch It Peak in Night Skies
35	Covid Protection	Seth Wernig/Associated Press	How to Double Mask Correctly	How NOT to Wear a Mask
36	Israel	How to Disinfect Your Space on an Airplane	Kushner's Absurd Peace Plan Has Failed	Trump's Middle East Peace Plan Exposes the Ugly Truth
37	Church	How the Palestinian-Israeli Peace Process Became a Farce	In Shift for Church, Pope Francis Voices Support for Same-Sex Civil Unions	United Methodists Tighten Ban on Same-Sex Marriage and Gay Clergy
38	Race Farms	Vatican Says Priests Can't Bless Same-Sex Unions	Country House Wins Kentucky Derby After Maximum Security Is Disqualified	Expert Picks: Who Will Win the Kentucky Derby?
39	Animals	How Maximum Security Was Disqualified From the Kentucky Derby	Whisky Knows Her Toys	A New Viral Outbreak Is Killing Rabbits
40	Judaism	The Dog Park Is Bad, Actually	'Plague on a Biblical Scale': Hasidic Families Hit Hard by Virus	\$15,000 Fine After Secret Hasidic Wedding Draws Thousands of Guests

C GPT Polarization Analysis

C.1 Article-level Polarization Measures

We used GPT-3.5-turbo, a large language model, to generate polarization scores for each article. That task was mostly infeasible with surveys, but the LLM can easily analyze the large corpus of 13508 articles and extract polarization scores for each of them.

We accessed a version of the model as of September 17th, 2023 through OpenAI API. To make the model comply with the task, we prompt it with the assignment prior to feeding it with article text (following -cite-). GPT-3.5-turbo model has been optimized for chat using the Chat Completions API, which means that it generates a response based on the provided chat history as input. Therefore, we use the same exact prompt as chat history input every time we run a model for a different article. We prompt the model as follows:

User: "I will provide you with a text from a news article. Please rate from 1 to 5 how polarizing this article is."

Model: "Sure, please provide the text of a news article and I'll do my best to rate the polarization level of each news article based on a scale from 1 to 5, where 1 means not polarizing at all, and 5 indicates a highly polarizing topic."

*User: *article text**

We ran the model for each article, seeding it with the conversation provided above and following with the article text. Note that we do not provide the model with the definition of polarization and rely on its own interpretation. The GPT model is stochastic and able to generate different responses with each iteration. We set the "temperature" argument of the model that controls the stochasticity of the output at a minimal value (0) to get more consistent answers by running the model only once for each article. In addition, each article text was cut to 1000 words to comply with the maximum number of tokens that GPT-3.5-turbo can process at once and speed up the process (the median number of words in an article is 1206, so it mostly affects larger articles). Table A4 provides examples of a few articles with different polarization levels, the GPT response after receiving these articles, and a score from 1 to 5 extracted from GPT response.

The GPT model provides a polarization score on a scale from 1 to 5. To be consistent with the way we treat the results of the survey, we standardized scores provided by GPT using z-scoring.

C.2 Topic-level Polarization Measures

In addition to generating polarization scores for each article, we used the model to generate polarization scores for LDA topics. We accessed a version of the model as of August 18th, 2023, and ran two types of questions. First, replicating the survey, we ask it to provide polarizing scores for each topic based only on the topic names. We use the same technique as above by promoting the prompting of the model as follows:

User: "I will provide you with a list of topics. Please rate them from 1 to 5 based on how polarizing they are."

Model: "Certainly! Please provide me with the list of topics, and I'll rate them on a scale of 1 to 5, with 1 being least polarizing and 5 being highly polarizing."

After that, we provide the model with an enumerated list of topic names. Even though we provide all topic

names together, the prompt is needed to run the model multiple times. We set the “temperature” argument of the model that controls the stochasticity of the output at a value of 1 and ran our prompt 10 times for each study to make our estimate more precise. For consistency of results, at each iteration, we seeded the model with exactly the same conversation described above. To be consistent with the way we treat the results of the survey, we standardized scores provided by GPT using z-scoring for each iteration (demeaned and divided by standard deviation for each iteration).

C.3 Keyword-level Polarization Measures

We were concerned that the LLM may be very sensitive to the particular way we named the topics. For robustness, we ran the second study where we provided a model with only the top 10 keywords for each topic and no topic name as in Table 1. The prompt is as follows:

User: “I will provide you with a list of keywords for topics. Please give the name for each topic and rate them from 1 to 5 based on how polarizing they are.”

Model: “Sure! Please provide me with the list of keywords for topics and I’ll be happy to give you the names for each topic and rate their level of polarization from 1 to 5, with 1 being the least polarizing and 5 being highly polarizing.”

Table A4: Examples of GPT-3.5 Responses for Article-level Polarization

Headline	Abstract	GPT-3.5-turbo Response	Score
Mary Oliver on Grief and Loss	Oliver, the hugely popular poet, died Thursday. Readers turned to her work to find comfort. Here's a selection of some of her best-known writing on loss and mourning.	Based on the provided text, I would rate the polarization level of this article as a 1. This article does not seem to contain any controversial or divisive elements. It primarily focuses on the life and work of Mary Oliver, highlighting her popularity and the impact of her poetry on readers. The article seems to be more informative and celebratory rather than polarizing.	1
Is This the Greatest Photo in Jazz History?	A quiet Sunday night in 1953. The Dodgers had just won the pennant. J.F.K. and Jacqueline Bouvier had just married. And four titans of bebop came together in a dive bar for a rare jam session.	Based on the provided text, I would rate the polarization level of this article as 1 out of 5. The article focuses on a specific event in jazz history, highlighting the performance of renowned musicians and the significance of a particular photograph. It does not touch upon controversial or divisive topics, making it relatively non-polarizing.	1
N.Y. Will Impose Quarantine on Visitors From States With Big Outbreaks	New Jersey and Connecticut will also require visitors to quarantine for two weeks. The rule reflects a stark shift in the course of the outbreak.	Based on the provided text, I would rate the polarization level of this article as a 2. While the article discusses the measures taken by New York and other states to control the spread of the coronavirus, it does not appear to present a highly polarizing viewpoint or controversial opinions. The focus is primarily on the actions taken by different states and the impact of the virus on public health.	2
Billie Eilish Sweeps Top Awards at the Grammys	The 18-year-old won five Grammys, including record of the year and best new artist, at a ceremony in which Kobe Bryant, who died on Sunday, was mourned.	I would rate this article as a 2 on the polarization scale. While the article covers various topics such as the Grammy Awards, Billie Eilish's success, Kobe Bryant's death, and controversies surrounding the Recording Academy, it does not seem to present any highly divisive or controversial viewpoints that would significantly polarize readers.	2
Small Businesses Wait for Cash as Disaster Loan Program Unravels	Owners were supposed to be able to get up to \$2 million. Now they're being told the cap is \$15,000 - if they can get any answers at all.	Based on the provided text, I would rate this article as a 3 on the polarization scale. While the article discusses the challenges faced by small businesses in accessing federal disaster loans and highlights the frustrations of applicants, it does not appear to take a strong stance or present a highly controversial viewpoint. The focus is primarily on the difficulties faced by business owners and the shortcomings of the loan program.	3

Continued on next page

Table A4: GPT Response Examples

Headline	Abstract	GPT-3.5-turbo Response	Score
4 Years After an Execution, a Different Man's DNA Is Found on the Murder Weapon	Lawyers' request to conduct additional DNA testing before Ledell Lee was executed had been denied.	Based on the provided text, I would rate this article as a 3 on the polarization scale. While the article discusses a controversial topic, namely the potential wrongful conviction and execution of Ledell Lee, it does not heavily emphasize or focus on polarizing viewpoints or opinions. The article primarily presents factual information, statements from involved parties, and updates on the case.	3
Our Democracy's Near-Death Experience	Now is no time for complacency. The next Congress must shore up our institutions.	Based on the provided text, I would rate this article as a 4 on the polarization scale. The article discusses the efforts made by President Trump and his allies to undermine the democratic process and challenges the legitimacy of his claims. It also emphasizes the importance of upcoming elections in Georgia and suggests that the actions of Republican enablers could subvert the democratic process. While the article presents a critical perspective on President Trump and his supporters, it does not explicitly attack or demonize them. However, the topic itself is inherently polarizing, as it involves discussions of election fraud, partisan politics, and the future of democracy.	4
I Don't Want to Be the Strong Female Lead	In pop culture, women are often objectified and disposed of. But even when we're not being victimized, the alternative leaves much to be desired.	Based on the provided text, I would rate this article as a 4 on the polarization scale. The article discusses polarizing topics such as gender roles, objectification of women, violence against women, and the portrayal of female characters in media. These topics often elicit strong and opposing opinions, leading to polarization among readers.	4
In Trump I Trust	The president's political project requires more than a few leaps of faith.	Based on the provided text, I would rate this article as a 5 in terms of polarization. The article presents a series of statements that are highly polarizing and divisive, expressing extreme beliefs and opinions about various political topics. The language used is strongly biased and aimed at promoting a specific viewpoint, which can contribute to further polarization among readers.	5
ISIS Enshrines a Theology of Rape	Claiming the Quran's support, the Islamic State codifies sex slavery in conquered regions of Iraq and Syria and uses the practice as a recruiting tool.	Based on the provided text, I would rate this article as a 5 in terms of polarization. The article discusses the horrific acts committed by Islamic State fighters, specifically focusing on the systematic rape of women and girls from the Yazidi religious minority. This topic is highly polarizing as it involves religious extremism, violence, and the violation of human rights. It is likely to evoke strong emotional responses and differing opinions on the subject matter.	5

D Appendix for Survey Analysis

D.1 Survey Questionnaire

We start the survey with the following preamble:

“We are interested in your views on news coverage and political polarization in mainstream media. There are no right or wrong answers; we are simply interested in your opinions.

Your responses are anonymous; we do not know your identity or have any ability to track it.”

After the preamble, participants were asked to rate how politically polarizing 10 randomly selected topics are (which are presented in random order). Figure A7 provides an example of such a question.

Please read the following definition of **Political Polarization**: A news topic is considered politically polarized if Democrats and Republicans hold strongly opposing views on the topic.

Now, rate how politically polarized the news coverage on the following topics is:

	Not at all polarized				Extremely polarized
Health Research, Lifestyle Advice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Israel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Emotions and Feelings	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Judaism	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sports	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
New York City	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nature	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Donald Trump	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Joe Biden	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Real Estate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Figure A7: Polarization question

At the end of the survey, participants were also asked to answer a few questions on their demographics and news reading and sharing habits, as shown below. Topics in question 4 correspond to the topics presented in the main question above and are shown in the same order (which was randomized) as Figure A7. The options in questions 6 and 8 were randomized for each participant to avoid order effects.

“You have finished the main part of the survey. Please tell us a little bit about yourself and your news reading habits. Note that all data are anonymized, and answers to this survey cannot be traced back to the respondent to identify you.”

1. Age

2. Gender Identity

- Male
- Female
- Non-binary / third gender
- Prefer not to say

3. How would you classify your political affiliation?

- Republican
- Democrat
- Independent

Figure A8: Demographics questions

4. How important is to you that your social circle knows of your opinions on and where you stand on the following topics:

	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Health Research, Lifestyle Advice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Israel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Emotions and Feelings	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Judaism	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sports	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
New York City	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nature	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Donald Trump	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Joe Biden	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Real Estate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5. In the last 12 months, how often have you visited news websites and/or read newspapers?

- Multiple times a day
- About once a day
- 2-3 times a week
- 2-3 times a month
- Less than once a month

6. Choose the top three news sources you typically read or follow:

<input type="checkbox"/> The New York Times
<input type="checkbox"/> The New Yorker
<input type="checkbox"/> Reuters
<input type="checkbox"/> Fox News
<input type="checkbox"/> Other
<input type="checkbox"/> CNN
<input type="checkbox"/> BBC
<input type="checkbox"/> NBC
<input type="checkbox"/> Washington Post
<input type="checkbox"/> Wall Street Journal

7. If you read a news article that you find interesting, how likely are you to share it with others?

Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8. If you do share it, what medium are you likely to use:

<input type="checkbox"/> Email
<input type="checkbox"/> Facebook
<input type="checkbox"/> Twitter
<input type="checkbox"/> Other Social Media
<input type="checkbox"/> Private Message
<input type="checkbox"/> None of the above

Figure A9: Questions on news reading and sharing preferences

D.2 Details of Survey Analysis

Of the 193 respondents who completed the survey, 100 (51.8%) were male, 92 (47.7 %) female, and 1 non-binary. The median respondent is 20 years old (average age is 20.75 with a standard deviation of 1.95) and spent around 2 minutes filling out the survey. In terms of political affiliation, 107 respondents identify as democrats, 18 as republicans, and 68 as independent. 71% of them visited news websites and/or read newspapers at least 2-3 times per week in the last 12 months and only 8.3% read news less than once a month. On average respondents follow 3 news sources from the list¹⁰. The most popular sources were CNN and BBC (145 and 137), followed by Reuters, Washington Post, and Other (70, 61, and 53). 45.6% of the users are somewhat (36.3%) or extremely (9.3%) likely to share an article if they find it interesting. And among

¹⁰The New York Times, CNN, Reuters, Wall Street Journal, BBC, Fox News, NBC, Washington Post, The New Yorker, Other

these, 81.8% would choose to share through a private message, 47.7% other social media, 17% email, 8% Twitter, and 8% Facebook.

Table A5 summarizes the main findings from the survey. For each topic, it shows the number of respondents that answered questions for that topic, the average score for how politically polarized they consider that topic to be, and how important it is to them that their social circle knows of their opinions on that topic. Since some respondents can give consistently higher or lower ratings than others, both scores are standardized using z-scoring (i.e., demeaned for each respondent and divided by the standard deviation for that respondent) (Bojd and Yoganarasimhan, 2022). Topics are sorted by the average polarization score.

Table A5: Survey results

	No. of observations	Polarization score	Social Signaling Score
Politics	53.0	0.89	-0.05
Donald Trump	50.0	0.88	0.11
Elections	45.0	0.79	0.21
Joe Biden	43.0	0.78	-0.25
Political Investigations	51.0	0.77	-0.27
Black Lives Matter	55.0	0.72	0.55
Women’s Issues and Sexual Harassment	43.0	0.69	1.02
Racial Identity and History	47.0	0.57	0.83
Judicial System	61.0	0.57	-0.01
American Military	49.0	0.46	-0.16
Coronavirus Pandemic	42.0	0.45	0.77
Covid Vaccine	45.0	0.44	0.39
Power, Energy Supply, and Climate	52.0	0.38	0.39
Covid Protection	49.0	0.36	0.45
World News	41.0	0.34	0.12
Public Health and Medicine	36.0	0.33	0.36
Israel	48.0	0.33	-0.38
Christianity and Church	46.0	0.25	-0.32
China, India, International Travel	42.0	0.25	-0.34
Business	47.0	0.20	0.65
Russia	41.0	0.14	-0.08
Social Media	42.0	0.08	-0.24
Education, School System	45.0	-0.04	0.37
Science	46.0	-0.12	-0.03
Money and Personal Finance	43.0	-0.18	0.17
Judaism	50.0	-0.18	-0.70
Family	39.0	-0.32	0.74
New York City	42.0	-0.33	-0.73
Emotion and Feelings	43.0	-0.34	0.42
Nature	42.0	-0.45	-0.19
Health Research, Lifestyle Advice	39.0	-0.46	-0.01
Real Estate	58.0	-0.60	-0.37
Music/Movies	46.0	-0.62	0.29
Books	47.0	-0.82	-0.25
Sports	58.0	-0.90	-0.03
Horse Racing and Farms	44.0	-0.93	-1.10
Art, Planes	46.0	-0.95	-0.80
Food	42.0	-1.00	0.21
Architecture	50.0	-1.26	-0.84

Continued on next page

Table A5: Survey results (standardized)

	No. of observations	Polarization score	Social Signaling Score
Pets and Animals	52.0	-1.27	-0.33

E Appendix for validity checks

We now present a series of robustness checks for the results shown in Table 4. Recall that, for Model (1), we use the difference between an article's rank in the two lists as the outcome variable. In that specification, for articles that are ranked in one list but unranked in another list, we treat their rank in the unranked list as 25. We now examine whether our results are sensitive to this assumption by considering two alternative options – (a) Model (3), where we treat the rank of unranked articles as 30, and (b) Model (4), where we treat the rank of unranked articles as 35. As shown in Table A7 results from regressions where we consider this alternative measure of rank difference are qualitatively similar to those in Model (1) in the main text. In both cases, the results are qualitatively similar to those from Model (1) in the main text.

To address the concern that results could be driven by emotions, we consider a version of Model (1) with controls for the usage of emotional words in the text. We use Linguistic Inquiry and Word Count (LIWC) 2015 software to get controls for general positive and negative emotions, as well as specific emotions such as anxiety, anger, and sadness. The results are shown in Model (5) in Table A7. Notice that the estimates are almost the same magnitude (and significance) as in Model 1 in Table 4).

Finally, we consider a model where we only consider the polarization score as an explanatory variable and other control variables, without including the topics; see Model (6). As expected, we find that the coefficient of polarization is over-estimated. This happens because some topics tend to have more polarized reporting, on average, and these topics also tend to be shared more on Facebook. Thus, it is important to control for topics, since it allows us to estimate the impact of polarization within topic.

Table A6: Other Specifications

Dependent Variables:	Difference in ranks (E-F)			
	Unranked=30 (3) OLS	Unranked=35 (4) OLS	Emotions (5) OLS	No Topics (6) OLS
<i>Variables</i>				
Polarization score	1.459*** (0.140)	1.831*** (0.173)	1.089*** (0.110)	1.905*** (0.090)
Family	17.681*** (2.110)	22.578*** (2.597)	12.247*** (1.664)	
Politics	-4.626* (2.359)	-4.875* (2.888)	-4.356** (1.881)	
Emotions and Feelings	7.914*** (2.321)	10.287*** (2.851)	6.438*** (1.840)	
Coronavirus Pandemic	24.580*** (1.993)	30.807*** (2.449)	18.121*** (1.567)	
Books	-3.026 (2.567)	-3.393 (3.173)	-3.173 (1.999)	
Nature	2.230 (2.357)	2.443 (2.891)	1.699 (1.842)	
Women's Issues, Sexual Harassment	35.769*** (2.531)	44.475*** (3.133)	26.811*** (1.954)	
Business	-5.969* (3.094)	-7.598** (3.811)	-4.429* (2.403)	
Education, School System	4.330* (2.437)	5.588* (3.003)	2.956 (1.906)	
American Military	21.119*** (2.282)	26.263*** (2.805)	15.615*** (1.817)	
China, India, International Travel	31.763*** (2.897)	39.719*** (3.563)	23.848*** (2.259)	
Power, Energy Supply, and Climate	8.610*** (2.785)	11.041*** (3.456)	5.850*** (2.147)	
Judaism	-38.727*** (8.484)	-50.856*** (10.651)	-27.333*** (6.387)	
Architecture	-17.706*** (2.382)	-22.517*** (2.938)	-12.805*** (1.850)	

Continued on next page

Table A6: Other specifications

Dependent Variables: Model:	Difference in ranks (E-F)							
	Unranked=30 (3) OLS		Unranked=35 (4) OLS		Emotions (5) OLS	No Topics (6) OLS		
Money, Personal Finance	9.508***	(2.049)	12.006***	(2.517)	6.701***	(1.622)		
New York City	10.994***	(2.640)	14.318***	(3.257)	7.318***	(2.053)		
Music/Movies	14.819***	(2.726)	18.797***	(3.387)	10.989***	(2.091)		
Health Research, Lifestyle Advice	-4.616**	(2.252)	-5.252*	(2.756)	-4.154**	(1.779)		
Black Lives Matter	28.740***	(2.173)	35.904***	(2.646)	21.138***	(1.739)		
Donald Trump	19.691***	(2.406)	24.525***	(2.964)	14.970***	(1.880)		
Elections	17.653***	(2.146)	22.219***	(2.625)	12.768***	(1.698)		
Joe Biden	15.034***	(2.244)	18.996***	(2.761)	11.211***	(1.753)		
Political Investigations	20.301***	(1.994)	25.352***	(2.438)	15.083***	(1.577)		
Public Health and Medicine	10.162***	(2.330)	12.872***	(2.858)	7.316***	(1.865)		
Racial Identity and History	8.209***	(2.470)	10.311***	(3.062)	5.850***	(1.910)		
Supreme Court and Judicial System	14.248***	(2.388)	17.456***	(2.947)	10.898***	(1.854)		
Food	-2.685	(2.327)	-3.289	(2.862)	-2.084	(1.820)		
Covid Vaccine	23.020***	(2.304)	28.481***	(2.834)	17.798***	(1.801)		
Art, Planes	-7.681**	(3.271)	-9.608**	(4.033)	-6.087**	(2.535)		
Covid Protection	-3.409	(2.451)	-3.605	(3.028)	-2.871	(1.914)		
Christianity and Church	10.032***	(3.552)	13.127***	(4.406)	7.133***	(2.732)		
Horse Racing and Farms	8.942	(5.759)	10.892	(7.177)	6.699	(4.403)		
World News	2.294	(3.036)	2.545	(3.748)	1.818	(2.348)		
Russia	12.326***	(2.515)	15.125***	(3.089)	9.887***	(1.963)		
Real Estate	-9.132***	(2.949)	-10.950***	(3.612)	-7.506***	(2.328)		
Sports	10.614***	(3.411)	12.797***	(4.190)	8.112***	(2.674)		
Science	5.835*	(3.541)	7.219	(4.387)	4.269	(2.724)		
Pets and Animals	-27.987***	(6.093)	-34.195***	(7.564)	-22.081***	(4.707)		
Israel	7.935**	(3.771)	10.050**	(4.606)	5.601*	(2.945)		
Headline length	0.118***	(0.044)	0.149***	(0.054)	0.085**	(0.033)	0.247***	(0.035)
Snippet length (standardized)	0.706***	(0.122)	0.853***	(0.150)	0.554***	(0.095)	0.776***	(0.096)
Word Count (standardized)	-1.514***	(0.122)	-1.876***	(0.151)	-1.149***	(0.094)	-1.120***	(0.099)
In Print (binary)	-2.526***	(0.249)	-3.188***	(0.308)	-1.854***	(0.194)	-3.021***	(0.195)
(Intercept)	-30.428***	(2.005)	-38.196***	(2.462)	-22.505***	(1.583)	-13.147***	(0.444)
LIWC Affective					-1.699**	(0.805)		
LIWC Positive emotions					1.584*	(0.812)		
LIWC Negative emotions					1.496*	(0.828)		
LIWC Anxiety					-0.065	(0.328)		
LIWC Anger					0.355	(0.317)		
LIWC Sadness					1.148***	(0.331)		
<i>Controls</i>								
Days after release		Yes		Yes		Yes		Yes
<i>Fixed-effects</i>								
Section name		Yes		Yes		Yes		Yes
<i>Fit statistics</i>								
Observations		23,580		23,580		23,580		23,580
R ²		0.34288		0.34329		0.33654		0.26935
Adjusted R ²		0.34036		0.34077		0.33383		0.26777

Clustered (date_id) standard-errors in parentheses

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

F Change over Time

Table A7: Change Over Time

Dependent Variables: Model:	Difference in ranks (E-F) (6) OLS	
<i>Variables</i>		
Post elections	-1.509***	(0.138)
Post × Polarization score	-0.554***	(0.177)
Polarization score	1.132***	(0.118)
Family	12.485***	(1.657)
Politics	-4.185**	(1.884)
Emotions and Feelings	5.369***	(1.839)
Coronavirus Pandemic	17.876***	(1.567)
Books	-3.125	(2.005)
Nature	1.644	(1.870)
Women's Issues, Sexual Harassment	28.424***	(1.946)
Business	-4.342*	(2.433)
Education, School System	2.684	(1.920)
American Military	15.306***	(1.792)
China, India, International Travel	23.190***	(2.278)
Power, Energy Supply, and Climate	6.464***	(2.168)
Judaism	-28.518***	(6.738)
Architecture	-13.090***	(1.866)
Money, Personal Finance	6.366***	(1.613)
New York City	7.365***	(2.089)
Music/Movies	11.310***	(2.112)
Health Research, Lifestyle Advice	-4.469**	(1.789)
Black Lives Matter	21.642***	(1.742)
Donald Trump	14.029***	(1.922)
Elections	14.249***	(1.738)
Joe Biden	10.357***	(1.770)
Political Investigations	14.696***	(1.594)
Public Health and Medicine	7.017***	(1.842)
Racial Identity and History	5.822***	(1.911)
Supreme Court and Judicial System	10.591***	(1.853)
Food	-1.906	(1.833)
Covid Vaccine	19.449***	(1.855)
Art, Planes	-6.079**	(2.563)
Covid Protection	-3.007	(1.931)
Christianity and Church	6.415**	(2.738)
Horse Racing and Farms	6.722	(4.469)
World News	1.680	(2.381)
Russia	8.438***	(1.992)
Real Estate	-6.822***	(2.340)
Sports	8.619***	(2.690)
Science	4.310	(2.801)
Pets and Animals	-19.781***	(4.744)
Israel	5.556*	(2.882)
Headline length	0.105*** (0.034)	
Snippet length (standardized)	0.569*** (0.098)	
Word Count (standardized)	-1.133*** (0.095)	

Continued on next page

Table A7: Other specifications

Dependent Variables: Model:	Difference in ranks (E-F) (6) OLS
In print (binary)	-1.797*** (0.197)
(Intercept)	-21.247*** (1.578)
<i>Controls</i>	
Days after release	Yes
<i>Fixed-effects</i>	
Section name	Yes
<i>Fit statistics</i>	
Observations	22,778
R ²	0.33702
Adjusted R ²	0.33436

Clustered (date_id) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

References

- S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- B. Bojd and H. Yoganarasimhan. Star-cursed lovers: Role of popularity information in online dating. *Marketing Science*, 41(1):73–92, 2022.
- M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.