Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

The Manifestation of Affective Polarization on Social Media: A Cross-Platform Supervised Machine Learning Approach

Christian Staal Bruun Overgaard, Josephine Lukito, Kaiya Soorholtz

The Center for Media Engagement, The University of Texas at Austin
csbo@utexas.edu, jlukito@utexas.edu, kaiya.soorholtz@austin.utexas.edu

Abstract

This project explores how affective polarization, defined as hostility towards people's political adversaries, manifests on social media. Whereas prior attempts have relied on sentiment analysis and bag-of-word approaches, we use supervised machine learning to capture the nuances of affective polarization in text on social media. Specifically, we fine-tune BERT to build a classifier that identifies expressions of affective polarization in posts shared on Facebook or Twitter during the first six months of the COVID-19 pandemic (n = 8,603,695). Focusing on this context allows us to study how affective polarization evolved on social media as the COVID-19 issue went from unfamiliar to highly political. We explore the temporal dynamics of affective polarization on Facebook and Twitter using ARIMA models and an outlier analysis of the first few months of the pandemic. Further, we examine the interplay between affective polarization and virality across the two platforms. The findings have important implications for those seeking to (1) capture affective polarization in text, and (2) understand how affective polarization manifests on social media. These implications are discussed.

Introduction

Political polarization is rising around the world (Westwood et al. 2018), not least in the United States (Iyengar, Sood, and Lelkes 2012). Although several forms of polarization exist, affective polarization, or negative feelings toward one's political opponents, has garnered much scholarly attention over the past decade (Iyengar et al. 2019). Much work has emphasized how social media might contribute to affective polarization (Allcott et al. 2020; Settle 2018) but little is known about how this type of polarization manifests on these platforms and change over time. Recent advances have made some progress towards this end (Marchal 2021; Yarchi, Baden, and Kligler-Vilenchik 2021) but have relied on methodologies such as sentiment analysis and bag-of-words approaches, which have limited ability to capture the nuances of affective polarization at the mass level.

The current research addresses these limitations by using supervised machine learning to build a classifier that can detect expressions of affective polarization in social media (Facebook, Twitter) content. As a case, we focus on COVID-19 posts/tweets from the first half year of the pandemic (n = 8,603,695). The COVID-19 issue is unique because it provides a glimpse not only into how people talk about contentious issues on social media but also into how affective polarization changes as issues go from politically neutral to highly partisan. To build the classifier, we follow a three-step approach. First, we explicate the concept of affective polarization in social media texts. Second, we conduct a content analysis to label a random set of posts/tweets (n = 3,194). Third, using this labeled dataset, we fine-tune a version of BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018) to create—and validate—a classifier that can capture affective polarization in social media texts. This allows us to (1) more accurately capture affective polarization in social media content than has been done before, and (2) understand how affective polarization manifested on Facebook and Twitter as the COVID-19 pandemic evolved during the first six months of 2020. The findings shed light on changes in public opinion pertaining to COVID-19 and have implications for recent concerns about growing political polarization around the world, and how divisiveness may be amplified by social media networks.

Literature Review

Affective Polarization

Recent decades have seen a substantial rise in affective polarization, or people's tendency to feel negative toward their political adversaries (Iyengar, Sood, and Lelkes 2012). Although other forms of polarization exist (Gentzkow 2016; Lelkes 2016; Wilson, Parker, and Feinberg 2020), affective polarization, in particular, has led to concerns over its societal implications (Broockman, Kalla, and Westwood 2022; Hartman et al. 2022; Hetherington and Rudolph 2015), spawning a rich scientific literature seeking to understand its nature, causes, consequences (for a review, see: Iyengar et al. 2019), as well as potential interventions to curb it (Hartman et al. 2022). The potential consequences of affective polarization should not be underestimated: high levels of affective polarization can bias how people interpret new issues (Druckman et al. 2021), increase distrust in interpersonal networks (Iyengar et al. 2019), and are associated with an increased tendency to believe in congenial claims whether they are factually true or not (citation withheld). Belief polarization might be particularly likely to occur in

times of crisis (Scheffer et al. 2022), which is, ironically, exactly when humans would benefit from a shared understanding of reality to foster effective decision-making and collaboration. From a normative perspective, affective polarization is more problematic to the vitality of democracies than polarization over policy preferences. Well-functioning democracies do not depend on citizens agreeing with each other; they do, however, require that people who disagree are willing and able to listen to each other and consider different points of view (Overgaard et al. 2022).

Many studies have highlighted the polarizing effects of social media (Settle 2018; Bail 2021); nevertheless, little is known about how affective polarization manifests on these platforms. Understanding what polarization looks like on social media can provide context for situating social media's role in polarizing the public (Tucker et al. 2018). Yet to gain such an understanding, it is necessary to go beyond the methodologies traditionally used employed in polarization research, such as surveys (Iyengar, Sood, and Lelkes 2012) or experiments (Huddy and Yair 2021).

Capturing Expressions of Affective Polarization With Computational Approaches

Affective polarization is typically operationalized as negative feelings toward outparty members (Iyengar et al. 2019). The current project does not use self-reported attitudes. Instead, we leverage the advantages of big data and computational methods, including the ability to study actual behavior as they occur in naturalistic settings (van Atteveldt and Peng 2018), to examine these attitudes as they are expressed in written content that may change rapidly over time.

Some scholars have used computational methods to study political phenomena in relation to news media and social media, for example by applying natural language processing techniques to measure polarization (including affective polarization specifically) in social media content. Common approaches in this area often apply dictionary approaches (Simchon, Brady, and Van Bavel 2020) or unsupervised machine learning techniques like Wordfish procedures (Hart, Chinn, and Soroka 2020) to detect political polarization in news or social media content. Whereas these efforts have examined political polarization in a broad sense, computational scholars have recently sought to measure more specific forms of political polarization in text. Yarchi et al. (2021), for example, measured interactional, positional, and affective polarization in Israelian social media content, pointing out the need to distinguish operationally between theoretically distinct forms of political polarization. We agree that such distinctions are crucial. Yarchi and colleagues' operationalization of affective polarization relies on sentiment analysis and is, as the authors rightly point out, limited because it does not take into account whether that sentiment is expressed toward any particular group, which is a core part of the definition of affective polarization. Similarly, Mentzer et al. (2020) focused on sentiment in tweets that talked about specific U.S. senate candidates but used a dictionary approach, making it difficult to know if expressions of negativity were aimed toward particular people or simply co-occurred with words describing political entities. Similarly, Marchal (2021) examined the sentiment of social media posts between like-minded or cross-cutting political audiences to examine the implications of positive and negative sentiment in these kinds of exchanges.

While these projects have made important headway in more efficiently detecting polarization, more nuanced approaches are necessary to ensure this detection is done efficiently and effectively. The aforementioned approaches, for example, tend to reduce polarization or rely on an overly broad operationalization. For example, the bag-of-words models rely on counting the positive and negative words in a given document. These strategies have at least two important limitations. First, dictionary-based approaches to sentiment analysis often perform poorly (van Atteveldt, van der Velden, and Boukes 2021) and are, therefore, problematic to rely on without human validation. A key strength of the current research is that we validate our classifier by comparing it to the gold standard—a carefully constructed human-labeled dataset based on a reliable content analysis (van Atteveldt, van der Velden, and Boukes 2021).

Second, affective polarization is distinct from constructs like negative sentiment and issue polarization. In this vein, Yarchi et al. (2021, p. 115) noted that their classification pertaining to affective polarization, "remains crude and needs to be further developed," and Marchal (2021) suggested using various approaches, including supervised machine learning, in future research. In the current research, we address the limitations of past scholarship, using a supervised machine learning approach to build—and validate—a classifier that can effectively identify expressions of affective polarization, with its nuanced differences from negative sentiment and issue polarization, in social media content.

To best leverage computational methods for text classification, it is important to first conceptualize the concept one seeks to operationalize (van Atteveldt, van der Velden, and Boukes 2021). We start by explicitly defining affective polarization as it relates to written content. In survey research, affective polarization has been defined as dislike of, or feelings of negativity toward, one's outparty members (Iyengar et al. 2019). We take a broader approach to conceptualize affective polarization in text, defining it as *expressions of dislike or negativity toward those the author of a social media post disagrees with politically*. This might take a number of forms, including calling one's opponents dishonest (e.g., "The left is lying"), malicious (e.g., "Democrats want to destroy our country"), or unintelligent (e.g., "Trump supporters are stupid") or by using pejorative terms to refer to groups (e.g., "Fascists") or specific people (e.g., "Nasty Pelosi"). Importantly, this definition distinguishes affective polarization from constructs like negative sentiment and issue polarization. Negative sentiment is a much broader category, whereas affective polarization specifically pertains to expressions of negativity or dislike *toward people that the poster disagrees with politically*. In short, affective polarization is inherently negative but negativity does not necessarily constitute affective polarization. Affective polarization (e.g., talking negatively about a politician) is also distinct from issue polarization (e.g., disliking a specific bill). Whereas

Polarized	Non-Polarized
"#COVID-19 and #Trump Both Make You Sick."	"Senate's bill would be 'terrible' for N.Y."
" Democrats do not care about our country, only empowering themselves, and their corrupt party."	"This pandemic is devastating the Navajo Nation, which already struggled with poverty."
"@realDonaldTrump Liar is a simpleton and pathological liar."	"Sad, More than 100,000 small businesses shutter as pandemic lockdowns devastate the economy."

Table 1: Sample sentences from posts/tweets classified as expressing affective polarization or not.

issue polarization is a matter of people *disagreeing* about politics, affective polarization is a matter of people *disliking* those they disagree with. Some examples of posts that were classified as affective polarized and not affective polarized are provided in Table 1. It is worth noting that some of the table's "non-polarized" posts do express negative sentiment or issue polarization; these examples are included to emphasize how affective polarization differs from these constructs.

Our conceptualization of affective polarization (which focuses on people that the author of a social media post disagrees with politically) is broader than the definition used in most survey research (which focuses on people's outparty members). This is useful for several reasons. Most importantly, Americans have become increasingly likely to express dislike of inparty members as the parties have fragmented. On the Democratic side, a sense of hostility has emerged between moderate liberals (who may support moderate politicians like President Biden) and stronger liberals (who may support politicians like Bernie Sanders or Alexandria Ocasio-Cortez) (Oliphant 2020). Among Republicans, the party has experienced a split between those who support former President Trump and those who oppose him (Kantor 2020). When it comes to expressions of affective polarization on social media, it is important to capture negative or hateful rhetoric aimed at those users disagree with politically—regardless of whether they normally support the same party. If a social media user posted something negative about President Biden, RINOs (Republican in Name Only), or people who refused to wear face masks, we want to capture it, regardless of which party the user tends to identify with. After all, polarized content might be problematic from a societal standpoint, regardless of who posted it.

Further, by focusing broadly on those the author of social media posts disagrees with, we ensure that our work can serve as a foundation for capturing affective polarization not only in the United States but also in multi-party systems. Finally, by taking this approach, we circumvent the need to infer the partisanship of social media users. Even though there are tools available for doing so, these tools come with limitations. Classifying social media users' partisanship, although possible, will inevitably result in errors. By avoiding making assumptions about the posters' partisanship, we, therefore, eliminate one potential source of errors.

A more general advantage of using computational approaches to detect affective polarization is the ability to conduct a longitudinal analysis in a more cost-efficient way. We take advantage of this to study how polarization may change over time across social media platforms and to identify key moments wherein affective polarization increases dramatically. This focus on tracking affective polarization as it manifests on social media over time sets our work apart from prior scholarship that has used computational methods to study affective polarization on social media.

In the current study, we focus on Twitter and Facebook to capture content on some of the most influential social media platforms in the U.S. Despite being the most widely used social media platform in America (Clement 2020), Facebook is relatively understudied compared to Twitter (Matamoros-Fernández and Farkas 2021; Tucker et al. 2018). Yet, Twitter is also important to consider in the current project, especially given the controversies that arose over former President Donald J. Trump using pejorative terms (e.g., "China Virus") to describe the pandemic in some of his tweets in March 2020[1]. Our multi-platform focus addresses an important gap in the polarization literature, which tends to lack cross- and multi-platform studies (Tucker et al. 2018). Including both Twitter and Facebook in our study allows us to not only study affective polarization on each platform but also examine whether affective polarization on either platform on a given day tends to be predictive of polarization on the other platform the following day.

Only a handful of studies have compared levels of affective polarization across platforms; for example, previous literature has found greater levels of cross-cutting political interactions on Facebook, as opposed to Twitter, but that disagreements (as a measure of affective polarization) are more pronounced on Twitter (Yarchi, Baden, and Kligler-Vilenchik 2021). Rather than conceptualizing social media as one unified entity within a model, as is often the case (e.g., (Törnberg et al. 2021)), we instead compare the two platforms. Given the limited scholarship comparing affective polarization content across platforms, we ask the following research question:

RQ1: Which proportion of political COVID-19-related texts on (a) Facebook and (b) Twitter express affective polarization?

Affective Polarization and Virality

Although little research has examined how affective polarization relates to virality (i.e., how much a piece of content spreads, for example, by being shared or retweeted), there are some indications that polarized voices are amplified more than less polarized voices. For example, Hong and Kim (2016) found that moderate members of the U.S. House of Representatives have fewer followers on Twitter than House members with more extreme ideologies. Tweets from

[1]See the supplementary materials (S6)

political elites can also be particularly viral when they mention the poster's outparty, which tends to be framed negatively (Rathje, Bavel, and Linden 2021). As humans are generally biased towards attending to negative cues compared to positive cues, we may also expect that affectively polarized content may be more likely to "go viral" (Baumeister et al. 2001; Moore-Berg, Hameiri, and Bruneau 2020). Correspondingly, news consumers tend to be more drawn toward negative than positive news stories (Soroka and McAdams 2015; Trussler and Soroka 2014).

Additionally, affective polarization tends to manifest as emotionally laden content, and these moral or emotional foundations can drive the popularity of polarized content on social media (Arora et al. 2022). Previous scholarship has shown that networked platforms such as Twitter may be prone to both emotion and heightened content distribution through retweets (Myers and Leskovec 2014). Twitter content with negative emotions, such as anger and fear, is particularly retweet-worthy (Nanath and Joy 2023). While there is less literature on virality on Facebook, evidence suggests that negative and emotional content also receives more shares (Alhabash and McAlister 2015).

However, there may be moments during a social or political event where affectively polarized discourse increases. For example, developments in news stories can translate to more affective polarization on social media. We describe temporal moments when affectively polarized content suddenly and substantially increases as outliers. In doing so, we complement the literature reviewed, which examined affective polarization on social media platforms but did not track the phenomena over time. Owing to the sensitivity of social media discourse (i.e., that conversations online can change as a result of offline phenomena (Myers and Leskovec 2014), social media analysis has become popular for detecting when an outlier or abnormal event has happened (Chae et al. 2012; Blázquez-García et al. 2021). The present study builds on this literature to examine and understand when affective polarization outlier moments occur.

RQ2: When do temporal outliers of affective polarization surface on Facebook and Twitter?

Another explanation for why affective polarization may increase on a platform is its relationship to another platform. In other words, trends on one platform may impact discourse on another. This idea is suggested in cross-media agenda-setting theory, which proposes that content on one platform may set the topical agenda of another platform. Notably, social media has been shown to set the agenda for news (Gilardi et al. 2022a). The ability for content on one platform to set the agenda on another platform is most commonly studied in relation to misinformation and disinformation (e.g., (Pierri, Artoni, and Ceri 2020; Ginossar et al. 2022)). However, it is also possible that increases in affective polarization on one platform may correlate with increases in affective polarization elsewhere.

RQ3: Can affective polarization on Facebook or Twitter be used to predict affective polarization on the other platform the next day?

RQ4: Are moments of increased affective polarization prone to increased virality?

The COVID-19 Pandemic and Polarization

In this study, we examine affective polarization in political discourse about the COVID-19 pandemic on social media. In doing so, our work complements the study of similar constructs like hate speech, sentiment, and emotions on social media during the COVID-19 pandemic (Fan, Yu, and Yin 2020; Hu et al. 2021; Jang et al. 2021; Tsao et al. 2021). The COVID-19 pandemic presents a unique opportunity to study not only how people talk about controversial issues but also how affective polarization arises over time as issues become politicized. This opportunity presents itself because COVID-19 emerged as a new issue in early 2020, which few people knew anything about or had political attitudes towards, and then grew into a very contentious political issue in the span of a few months. Research points to high degrees of issue polarization among Republicans and Democrats over the COVID-19 issue (Gallup 2020; Pew Research Center 2020), which also affected their behavior during the pandemic (Gollwitzer et al. 2020). Nascent research regarding the political communication surrounding COVID-19 notes that the issue was politicized and polarized quickly, among both elites (Green et al. 2020) and in news coverage (Hart, Chinn, and Soroka 2020). It stands to reason that virality as an amplification mechanism on social media may help make affective polarization messages on social media gain even more popularity or engagement.

Method

Data

Facebook Data The Facebook data was collected using the CrowdTangle (2020) platform, a popular social media monitoring tool used by political communication scholars (Frischlich 2020; Larsson 2020). A search was made for COVID-19-related posts, which had been posted in English-speaking Facebook groups from January 21 to May 31, 2020. As search terms, we used common words related to the pandemic (e.g., coronavirus, covid, and covid-19), alternative spellings (e.g., corona virus), generic terms (e.g., pandemic), slang (e.g., missrona, covidiot), and pejorative terms (e.g., kung flu, wuhan virus). CrowdTangle's search is not case-sensitive. Text written in attached images is searchable with CrowdTangle and was treated as regular text. Posts that mentioned one or more search terms in their URLs but not in their main text or images were excluded. Facebook groups, rather than Facebook pages, were chosen as the focus of this study because many of the Facebook pages sharing posts about COVID-19 were news organizations. Ideally, the study would have included posts shared by Facebook users on their walls; however, this data, which comes with ethical and privacy concerns, was not available. After removing duplicate posts, the dataset consisted of 4,445,858 Facebook posts.

Because this project focused specifically on U.S. politics, we built on prior research (Hart, Chinn, and Soroka 2020; Simchon, Brady, and Van Bavel 2020) to create a dictionary that could identify posts that mentioned words related to U.S. politics. This dictionary included general words about U.S. politics (e.g., Democrats), and the names of party lead-

ers (e.g., Trump), prominent presidential candidates (e.g., Biden), and the congress members with the most followers on Twitter and Facebook (e.g., Mitt Romney). Given our specific focus on the U.S. context, generic political words (e.g., politics, president) were left out. Due to our focus on political polarization, we also included frequently used condescending nicknames (e.g., Sleepy Joe). For very prominent politicians who are often referred to by only their last name (e.g., Biden, Pelosi, Trump), we included just their last name; for politicians who were less prominent or had last names that could easily be referring to other people (e.g., Joe Kennedy, Rand Paul), their full names where used. After applying the politics dictionary, the final dataset consisted of 274,849 posts.

Twitter Data Twitter data was collected, for the same period, using the Twitter API for Academic Research. This was done in R, using the academictwitteR package. We searched for tweets that included at least one word from our COVID-19 dictionary and at least one word from our politics dictionary, and the query excluded retweets and non-English tweets. After removing duplicates, this yielded a dataset consisting of 8,328,846 tweets.

Supervised Machine Learning

To measure affective polarization, we created a supervised machine learning classifier. Two research assistants hand-coded a sample of 3,194 randomly selected texts to denote whether each text expressed affective polarization (n = 854) or not (n = 2,340).[2] Stratification was used to ensure that 80% of the messages in the sample were tweets whereas 20% were Facebook posts (see Grimmer and Stewart 2013). Before hand-coding this sample, the two coders had been trained until they reached intercoder reliability (92.7% agreement; Krippendorff's alpha = 0.774). The training and hand-coding were based on a formal codebook that operationalized affective polarization (see below).

With this sample, we built a classifier to label whether each post or tweet in our dataset expressed affective polarization. At first, we ran a series of Bag-of-Words models (KNN, SVM, random forests, logistic regression, Naïve Bayes, Xgboost). However, none of these models performed well, with F-scores consistently falling below 0.4. Capturing the complexity of affective polarization in text, it seems, is difficult with Bag-of-Word approaches, even when trying a variety of approaches. This echoes our initial suspicion that the Bag-of-Words approaches employed in prior attempts to capture affective polarization in text might be subject to serious limitations.

Therefore, we created a fine-tuned version of BERT (Devlin et al. 2018). This was implemented in Python using the PyTorch and Transformers packages.[3] as a starting point and then fine-tuning (i.e., training) it on our labeled dataset.[4] The fine-tuning was done by splitting the labeled dataset into a training (n = 2,555), validation (n = 319), and test set (n = 320). The training and validation sets were used to construct the fine-tuned version of the model, whereas the test set was used to evaluate the model's performance. The model evaluation, which was done using the Scikit-learn package for Python, showed a macro precision of .72 (weighted average = .82), a macro recall of .76 (weighted average = .82), a macro F-score of .77 (weighted average = .82) (AUC = .78).[5] The precision metric refers to the proportion of posts identified as expressing affective polarization that truly did express affective polarization; the recall metric refers to the proportion of posts actually expressing affective polarization that was correctly identified as expressing affective polarization. The F-score is akin to a harmonic mean that is calculated from the precision and recall; it can range from zero to one and is an overall metric used to determine model performance. An F-score of about 0.80 is on par with most political communication research using supervised machine learning (Das et al. 2021; Haworth et al. 2021; Matalon et al. 2021). As for computational research on affective polarization in social media users, this constitutes a crucial improvement over prior projects, which have not validated inferences against the gold standard—a labeled dataset based on a reliable content analysis (van Atteveldt, van der Velden, and Boukes 2021).

The full list of terms queried in the data collection, the dictionary used to determine whether a message was political, and the affective polarization codebook used to build the labeled data are included in the supplemental materials, as are the classifiers' performance metrics and the final model's confusion matrix.[6]

Time Series

Time series modeling was used to study the temporal relationship between affective polarization and virality across the two platforms. Time series modeling is a longitudinal technique that treats each dataset as a discrete and equidistant sequence of data points such that time t precedes time $t+1$. Owing to the rich temporal meta-data in digital media content (Giachanou and Crestani 2016; Ikeda et al. 2013), the use of time series modeling in computational social sciences has become increasingly popular (Wells et al. 2019).

To do this modeling, we first constructed four time series variables: (1) a daily proportion of affectively polarized posts on Facebook, (2) a daily proportion of affectively polarized tweets, (3) a daily count of retweets of political COVID-19 tweets, and (4) a daily count of shares of political COVID-19 Facebook posts.

Before proceeding with a multivariate analysis, we first wanted to understand each time series using two analyses: a time series outlier analysis using the "tsoutliers" package

[2]Of these, 599 were coded by both research assistants, and the remaining posts were divided between them. Disagreements within the initial batch of 599 were resolved by the authors.

[3]We used "bert-base-uncased" from the Transformers package.

[4]For tokenization, BertTokenizer from the Transformers package was used. The model ran for four epochs (learning rate of 1e-5).

[5]Accuracy = .82. Correlation between manual annotations and model predictions: r = .92, p <.001.

[6]The supplemental materials, as well as the manually labeled data and code used to train the classifier (license: CC-BY), can be accessed at https://t.ly/qX2AL, https://tinyurl.com/4dn6m3sj or osf.io/h5afe/?view_only=dba1ad98f9da4738907a7a91cea3c60b.

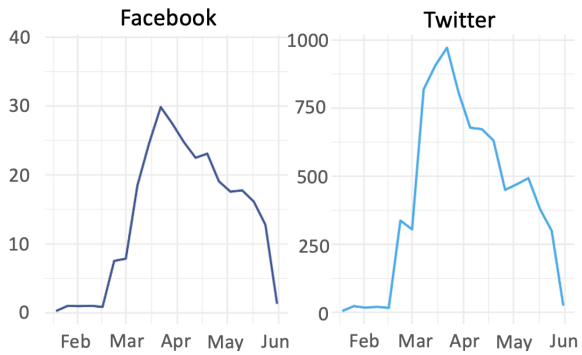


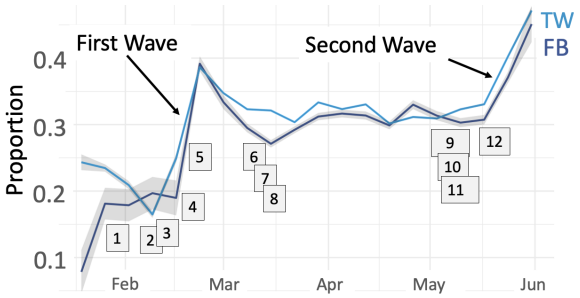
Figure 1: Weekly number of COVID-19 related posts or tweets (in thousands). (Note: The Facebook corpus only includes content posted in English-speaking Facebook groups. The two y-axes are scaled differently because there was a greater number of tweets than Facebook posts.)

(López-de Lacalle 2019) and four univariate ARIMA models (Pack 1990). ARIMA models are used to analyze three time series data-generating processes: mid-term autoregressive (AR) processes, long-term integrated (I) processes, and short-term moving average (MA) processes. ARIMA models are commonly used to forecast future time series points (Samaras, García-Barriocanal, and Sicilia 2020). Building on ARIMA modeling's prediction capabilities, time series outlier detection will find temporal moments where the data deviate from the predicted trend. In other words, when ARIMA modeling cannot predict the next data point because it is unexpectedly high or low, that point is deemed an outlier.

Following the aforementioned descriptive analysis, we then set to testing our research question by constructing a Vector Autoregression (VAR) model, a popular multi-variate time series model for assessing bi-directional relationships across multiple endogenous variables (Sims 1986; Gilardi et al. 2022b). Similar to ARIMA models, VAR models are also used in time series forecasting but have the advantage of predicting across multiple endogenous time series (Liu, Tseng, and Tseng 2018). In a VAR model, each time series is modeled as a function of the lagged variables for all other time series in the equation, making interpretation of partial-model results difficult. As a result, Granger Causality tests are commonly used to focus on the potential predictive power of specific variables within the VAR model (Toda and Phillips 1994). Thus, our analysis will focus on the Granger Causality tests.

Results

Before digging into the modeling, let us first explore the weekly volume of political COVID-19-related posts and tweets, which changed throughout the period (Figure 1). On Facebook and on Twitter, discussions about COVID-19 started gaining serious traction toward the end of February, rising rapidly until the third week of March. After March, the weekly volume of posts steadily declined throughout the rest of the period.



1 (1/30): First US COVID transmission; 2 (2/7): COVID kills Dr. Li Wenliang; 3 (2/10): COVID deaths > SARS deaths, globally; 4 (2/18): Diamond Princess quarantine; 5 (2/23): Italy lockdown; 6 (3/10): Trump uses pejorative terms like 'Wuhan virus'; 7 (3/11): WHO declares pandemic; 8 (3/13): Trump declares national emergency; 9 (5/1–5/10): US protests & mask controversy; 10 (5/6): NYC subway shuts down; 11 (5/7–5/10): White house staffers infected; 12 (5/18): Trump / hydroxychloroquine claims.

Figure 2: Affective polarization on Facebook (FB) and Twitter (TW). (Note: The Y-axis denotes the proportion of posts/tweets classified as polarized, at the weekly level. The grey shade areas indicate 95% confidence intervals.)

On the whole, the levels of affective polarization were similar across the two platforms, with 30.9% of the Facebook posts and 32.5% of the tweets being classified as affectively polarized. Examples of polarized and non-polarized texts in the dataset are provided in Table 1.

Changes in the proportion of affective polarization over time are displayed in Figure 2. On Twitter, the proportion of posts expressing affective polarization started at about 25% in late January 2020; on Facebook, it started close to zero. Towards mid-February, the lines for the two platforms converged at around 20%. After that, the trend is similar across the two platforms, occurring over two distinct waves. The first wave occurs in mid to late February, where the proportion rises from about 20% to about 40%. In March, the proportion settles at around 30% where it stays until May. At the end of May, the second wave of affective polarization occurs, with the proportion rising from 30% to about 45%.

The virality (shares, retweets) of the content also changed over time. As seen in Figure 3, virality on Twitter and Facebook follow similar trajectories, with a handful of exceptions; most notably, virality (measured as weekly shares or retweets per text) substantively, but temporarily, increase in late February, mid-April, and end of May.

Time Series Modeling

Results of the descriptive, univariate analysis suggest that the proportion of affective polarization posts on Twitter and Facebook follow similar processes, with the optimal models being an ARIMA(0,1,2) for the proportion of tweets (MAPE = 18.25), and an ARIMA(1,1,2) for the proportion of Facebook posts (MAPE = 15.52).[7] Notably, both

[7] An ARIMA model identifies three data-generating processes: (p, q, d). p, is a mid-term autoregressive process. q, is the long-term integration (most often measured as either 0 = no integration, also known as stationarity, or 1 = integration, suggesting non-

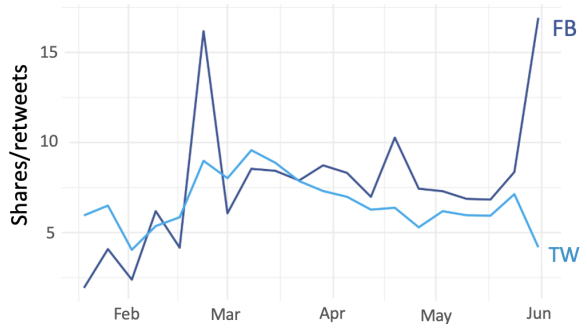


Figure 3: Virality on Facebook (FB) and Twitter (TW). (Note: The number of shares/retweets for each platform is aggregated at the weekly level.)

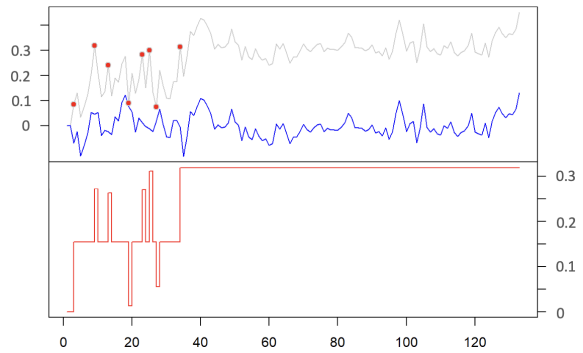


Figure 4: Outlier analysis for Facebook (Note: The blue line represents the data as predicted by the ARIMA model. The grey line represents the actual line data. Red dots indicate outliers. The red line below represents the type of outlier: additive outliers, transient change, or level shifts.)

univariate time series appear to be non-stationary (q = 1) and have two moving-average components. This confirms what we can visually see in Figure 2: affective polarization steadily, and permanently, increases over the timeframe of the analysis. The long-term data generating process, or the integrated component, was also confirmed using ADF tests for both Twitter (p = 0.02) and Facebook (p = 0.01), and ACF/PACF graphs, two common techniques for detecting non-stationarity (Noureen et al. 2019).

To further understand the univariate time series, we conducted outlier analyses to understand the spikes in affective polarization. Outlier analyses are used to detect data points in a time series that cannot be forecasted or predicted using an ARIMA model. There are three types of outliers: transient changes (outliers with a diminishing influence on the data-generating process), additive outliers (outliers that only last for one or two days), and level shifts (outliers that produce a permanent change in the data-generating process). Focusing on the proportion of affectively polarized tweets and Facebook posts, we find that all the outliers in these two variables occurred early in the time series, in the lead-up to the first wave of polarization as seen in Figures 4-5, with the Twitter time series yielding 3 outliers and the Facebook time series yielding 8. Curiously, on Twitter, affective polarization decreased first (in a temporary additive outlier and a separate transient change outlier) before having a positive additive outlier spike, suggesting that affective polarization was much lower in this early stage of the time series compared to the latter months. The volatility of affectively polarized discourse is especially noticeable on Facebook, though, as there are eight rapid positive and negative outliers, including two positive, long-term level shifts. [8]

Following the outlier analysis, we constructed a vector autoregression model to understand whether the proportions of affectively polarized tweets, the proportion of affectively polarized Facebook posts, the virality of political COVID-19 tweets, and the virality of political COVID-19 Facebook posts were temporally related. To prepare the data, we first-differenced the data to remove integrated components. Then, using the Bayesian Information Criterion (BIC) (Jones 2011), we determined an optimal lag structure of 1 (BIC = 3724.88). We constructed a VAR(1) model, treating all four time series as endogenous variables that could temporally predict one another. As VAR model results are typically presented as partial models (as each variable is measured as an equation of the lagged other variables), we supplemented this analysis with Granger causality tests, which are often used to test bivariate relationships within a VAR model (Uyheng and Carley 2020).

The Granger causality tests examining the relationship between the proportion of affectively polarized tweets and the virality of COVID-19 political tweets were not statistically significant in either direction. However, we find that the proportion of affectively polarized Facebook posts Granger caused virality of COVID-19 political Facebook posts (F = 5.33, p =.022) and vice versa (F = 5.23, p = .023). VAR results confirmed that this relationship was positive. In other words, a greater proportion of affectively polarized Facebook posts at time t could predict more viral political COVID-19 posts were on Facebook at time $t+1$, which (in turn) would increase the proportion of affectively polarized Facebook posts at the next time point.

Results of our Granger causality tests also suggest that the virality of political, COVID-19 tweets could predict the virality of political, COVID-19 Facebook posts at time $t+1$ (F = 7.51, p = 0.007), and vice versa (F = 7.36, p = 0.008). This unexpected result suggests a "diffusion of virality" from one platform to another.

Discussion

This research introduces and validates a new way of measuring expressions of affective polarization in social media texts. This was done by labeling a large number of texts (n = 3,194) according to a validated set of criteria. This set of texts was then used to build and validate a classifier by fine-

stationarity). d, is the short-term moving average process.

[8] Owing to how close the outliers are to one another, it is difficult to assess whether an external event may have contributed to any singular outlier. However, the "bunching" of outliers into a specific portion of the time series highlights the potentially unexpected rise in polarization in the early months of the pandemic.

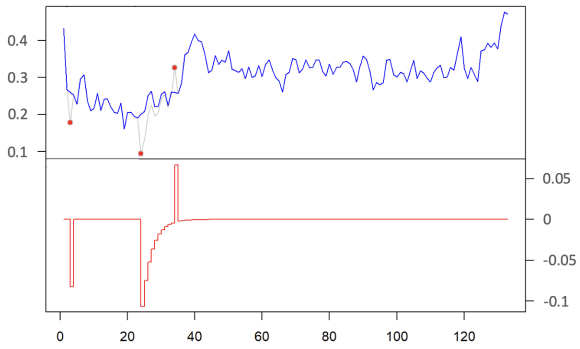


Figure 5: Outlier analysis for Twitter. (Note: The blue line represents the data as predicted by the ARIMA model and the grey line represents the actual line data, with red dots indicating the time point of the outlier. The red line below represents the type of outlier: additive outliers, transient change, or level shifts.)

tuning BERT (Devlin et al. 2018). Concentrating on content related to the COVID-19 pandemic, we then investigated how discourse around this topic polarized as the pandemic unfolded over the first two quarters of 2020. As part of this effort, we explored cross-platform dynamics as well as the interplay between virality and affective polarization. Our findings make several contributions to the polarization literature. Although past scholarship has used computational methods to measure affective polarization in social media content, our approach stands out methodologically (by pairing supervised machine learning with human validation) and theoretically (by tracking polarization and virality over time and across platforms). This allows us to offer a more comprehensive understanding of how affective polarization manifests and evolves on social media platforms. Further, by leveraging the context of the COVID-19 pandemic, we shed new light on how affective polarization arises over novel issues as they become increasingly political.

Measuring Affective Polarization in Text

Our most fundamental contribution is introducing and validating a classifier that can accurately capture affective polarization in social media texts. Prior efforts have grouped different types of polarization together or relied on overly reductive measures (e.g., dictionary approaches and/or sentiment analysis), limiting researchers' ability to capture the theoretical conceptualizations of these constructs (Marchal 2021; Mentzer et al. 2020; Rathje, Bavel, and Linden 2021; Yarchi, Baden, and Kligler-Vilenchik 2021).

Our first step was to explicate what it means for a social media text to express affective polarization, which we conceptualized as *expressions of dislike or negativity toward those the author of a social media post disagrees with about politics*. This definition builds on how affective polarization has been conceptualized in survey research, which has defined it as feelings of dislike or negativity aimed at the outparty (Iyengar, Sood, and Lelkes 2012; Iyengar et al. 2019). We take a broader approach, focusing on people the poster disagrees with, which we have argued has several advantages: (1) it better captures the recent fragmentation of both of the major U.S. political parties; (2) it allows the present research to serve as a building block for classifiers that function not only in the United States but also in multi-party systems; (3) it eliminates the need to make assumptions or inferences about the poster's partisanship, which, although possible, comes with limitations and errors.

Our conceptual definition of affective polarization serves as our foundation for building a classifier. This is important given that the main limitation of previous efforts (Marchal 2021; Mentzer et al. 2020; Rathje, Bavel, and Linden 2021; Yarchi, Baden, and Kligler-Vilenchik 2021) have been that their operationalizations have not clearly mapped onto the conceptual definition of affective polarization. Based on our conceptualization of affective polarization, we conducted a content analysis to capture this concept in tweets and Facebook posts. Two research assistants were trained based on a detailed codebook to reliably identify this concept in tweets and Facebook posts, and then labeled more than three thousand texts.

With this labeled dataset, we created a classifier, leveraging a fine-tuned version of BERT (Devlin et al. 2018), to identify affective polarization in social media texts at scale. This model performed substantially better than any of the Bag-of-Word approaches we tried, and it is fundamentally different from the attempts made by prior research to identify affective polarization in social media users at the mass level. Prior attempts have relied on dictionary-approach or sentiment analysis (Marchal 2021; Mentzer et al. 2020; Rathje, Bavel, and Linden 2021; Yarchi, Baden, and Kligler-Vilenchik 2021), which is problematic because these techniques have not been able to clearly distinguish affective polarization from negative sentiment or issue polarization. Affective polarization is indeed difficult to capture using automated approaches; we found that no Bag-of-Word strategies yielded F-scores greater than 0.40. Our classifier, which leverages BERT, performs substantially better, with an F-score of about 0.80. This speaks to a more general point: we were only able to assess the accuracy of the Bag-of-Word approaches because we had a gold standard (a dataset reliably labeled by human coders) to compare their performance with. We, therefore, echo van Atteveldt and colleagues' (2021) recommendation that automated text analyses must be validated by humans.

Understanding How Affective Polarization Arose Across Social Media Platforms As the COVID-19 Pandemic Unfolded

The current research also speaks to how affective polarization manifests on social media, particularly around unfamiliar issues that transform into hot-button topics. To this end, we focused on social media discourse about COVID-19 during the first six months of the pandemic in 2020. We find that affective polarization followed largely similar trajectories across Facebook and Twitter. On both platforms, affective polarization was relatively low in January and February, although at this point polarization was higher on Twitter

than on Facebook. Affective polarization then rose, following very similar trajectories across both platforms, in two waves. These similarities may suggest that we are observing trends that generalize to social media in a broad sense, and are not unique to one specific platform. Of course, without testing, it is impossible to know what these trends might look like on other platforms. We leave that as a task for future research and hope our classifier can aid such endeavors.

As for the observed trends, the first wave of polarization arose in mid-February, whereas the second wave arose toward the end of May. It is noteworthy that affective polarization seem to rise substantially across both Facebook and Twitter even before some of the controversies that arose over the issue, for example, when former President Donald J. Trump began using pejorative terms (e.g., "China Virus," "Wuhan Virus," and "Kung Flu") on Twitter in the first half or March 2020. This is surprising, especially in light of prior concerns and evidence that polarizing cues from political elites fuel mass polarization (Wilson, Parker, and Feinberg 2020; Bisgaard and Slothuus 2018). Our findings suggest that other mechanisms, besides cues from political elites, might have been more potent in fueling the animosity that arose over the COVID-19 issue. Furthermore, our findings fit with a possibility, raised by Iyengar et al. (2019), namely that affective polarization among regular citizens might influence political elites.

The first wave of polarization arose relatively early, even before a pandemic was declared by WHO, or a national emergency was declared in the U.S., in mid-March. Yet this wave unfolded at a time when COVID-19 came to be seen as a serious threat. In Early February, COVID-19's death toll in China surpassed the death toll of SARS during that crisis. On February 23, Italy imposed a lockdown, which was widely reported by international as well as U.S. news media. Although COVID-19 had yet to proliferate globally, media coverage made it clear that the disease might be both dangerous and difficult to stop from spreading. Some of this coverage included graphic depictions of COVID-19 victims (e.g., lying in hospital beds or body bags), which can exacerbate news consumers' feelings of anger and anxiety and even make them more prejudiced towards their outgroup members (Overgaard 2021). At this time, COVID-related anxiety was enough of a concern that the American Psychological Association published resources about coping with it. [9]

The second wave of polarization arose in May, as the total number of COVID-19 deaths passed 100,000. This wave co-occurred with some contentious political events, including lockdown protests (with controversies arising over masking), several White House staffers testing positive for COVID-19 (sparking criticism among Democrats), and President Trump saying he had personally been taking the controversial immunosuppressive drug hydroxychloroquine. This wave also occurred at a time when some people began experiencing lockdown fatigue and Zoom fatigue, and when major companies and universities announced that work would continue remotely for the foreseeable future.

Using time series analyses, we also identified the interplay between affective polarization and virality and the interplay between virality on the two platforms. First, we found that, on Facebook specifically, affective polarization on a given day is predictive of virality on the next day, and vice versa. However, we did not find a similar pattern on Twitter. This may suggest that affective polarization precedes viral engagement on Facebook, but not on Twitter. There may be several explanations for this, including potential differences in the user base (perhaps Facebook users are more likely to engage with affective polarization) or differences in the platform's infrastructure (such as the way in which algorithms on each respective platform treat retweets and shares).

Second, we find that virality on Facebook on a given day was predictive of virality on Twitter on the next day. This diffusion of virality from one platform to another suggests some linking of engagement across platforms-not necessarily causal, but temporally predictable. Perhaps popularity on one platform may motivate others to spread that content elsewhere. Or, given the high interest in reading COVID-19 information in the early months of the pandemic, the underlying content may have been viral regardless of the platform. Curiously, despite the reputation of Twitter as a platform that provides rapid, real-time information (Sakaki, Okazaki, and Matsuo 2010), it is actually Facebook virality that predicts Twitter virality, rather than the opposite. While unexpected, the COVID-19 pandemic persisted for far longer than an earthquake or tornado, which may have given time for affective polarization to flourish outside of Twitter.

These findings open up new avenues of scholarship on cross-platform virality and the spread of content–in this case, affectively polarized posts. Though the scholarship on affective polarization has not done enough to focus on cross-platform research and has, perhaps, done too little to describe how phenomena play out on social media before untangling their causal effects (Tucker et al. 2018), our findings suggest a need for more empirical work on affective polarization as it extends across platforms. One inherent limitation of this study is the generalizability of our findings to other platforms beyond Facebook and Twitter; we therefore encourage researchers to consider a broader range of platforms in the future.

Conclusion

This work also holds societal value. Affective polarization over COVID-19 can be problematic when making political or public health decisions (Gollwitzer et al. 2020), both at the societal level (support or opposition towards public health policies) and for individuals (individual health choices during a pandemic). A more nuanced view of this phenomenon, and how it comes about, can foster an understanding of how divisiveness arises in the age of social media, particularly around emergent political issues.

Although situated in a social media context, this project also contributes new knowledge about the changing nature of COVID-19 discourse and opinion formation. Public opinion is typically examined through surveys, which remained true for understanding public opinion about the COVID-19 pandemic. Social media data, as used in the current research, can add important nuances by complementing sur-

[9] For examples, see the supplemental materials (S6).

vey data in at least two important ways. First, social media data give an indication of people's actual behavior as opposed to their self-reported attitudes. Second, social media data makes it possible to track how phenomena unfold over time in much greater temporal detail (here, at the weekly level) than surveys would. Taken together, these advantages allow this study to add some nuances to what is known about how the nature of the COVID-19 issue changed over time, as a novel news story grew into a historic pandemic.

In sum, this paper introduced and validated a way of measuring affective polarization in text, which constitutes a substantial improvement over prior computational approaches. We further investigated how COVID-19 discourse on social media grew increasingly polarized as the pandemic unfolded, and we also shed new light on cross-platform dynamics and the interplay between polarization and virality. Besides adding theoretical insight to the growing literature on affective polarization, we believe these efforts can serve as a building block for future attempts to capture affective polarization in text. Given the complexities of affective polarization, we caution scholars to use careful human validation of automated computational ways of capturing this rather nuanced construct (for similar concerns, see: van Atteveldt, van der Velden, and Boukes 2021). By better understanding how polarization manifests on social media, we hope that scholars and digital architects will be better able to understand the interplay between polarization and social media, and better equipped to build constructive digital spaces.

Ethics Statement

Careful steps were taken to minimize ethical risks and potential negative outcomes. For data collection, we used publicly accessible data acquired through official APIs and archives. Our time series modeling is done in the aggregate, and our results speak to overall trends rather than the virality of any one message. We recognize that the outcome of our work may motivate or encourage affective polarization as a means for gaining virality. To minimize this risk, we focused on an ongoing public health issue that, by its nature, does not benefit from affective polarization. We believe this work may help identify affectively polarized messages on social media and may help public health experts identify potential moments of polarization before they gain traction. The authors have no conflicts of interest related to this project.

Acknowledgements

This work was supported by grants to the Center for Media Engagement by the William and Flora Hewlett Foundation and the John S. and James L. Knight Foundation. The authors would like to thank Oluwaseyi Odufuye and William Tran for assistance on the project. The authors would like to thank Natalie Jomini Stroud, Ashwin Rajadesingan, the Center for Media Engagement team, and four anonymous reviewers for their feedback.

References

Alhabash, S.; and McAlister, A. R. 2015. Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of Facebook and Twitter. *New media & society*, 17(8): 1317–1339.

Allcott, H.; Braghieri, L.; Eichmeyer, S.; and Gentzkow, M. 2020. The welfare effects of social media. *American Economic Review*, 110(3): 629–76.

Arora, S. D.; Singh, G. P.; Chakraborty, A.; and Maity, M. 2022. Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183: 121942.

Bail, C. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press. ISBN 978-0-691-21650-8.

Baumeister, R.; Bratslavsky, E.; Finkenauer, C.; and Vohs, K. 2001. Bad is stronger than good. *Review of General Psychology*, 5(4): 323–370.

Bisgaard, M.; and Slothuus, R. 2018. Partisan elites as culprits? How party cues shape partisan perceptual gaps. *American Journal of Political Science*, 62(2): 456–469.

Blázquez-García, A.; Conde, A.; Mori, U.; and Lozano, J. A. 2021. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3): 1–33.

Broockman, D. E.; Kalla, J. L.; and Westwood, S. J. 2022. Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not. *American Journal of Political Science*, 0(0): 1–21.

Chae, J.; Thom, D.; Bosch, H.; Jang, Y.; Maciejewski, R.; Ebert, D. S.; and Ertl, T. 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE conference on visual analytics science and technology (VAST)*, 143–152. IEEE.

Clement, J. 2020. Most popular social media apps in U.S.

Das, M.; Saha, P.; Dutt, R.; Goyal, P.; Mukherjee, A.; and Mathew, B. 2021. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 79–89.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Druckman, J. N.; Klar, S.; Krupnikov, Y.; Levendusky, M.; and Ryan, J. B. 2021. Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour*, 5(11): 28–38.

Fan, L.; Yu, H.; and Yin, Z. 2020. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. *Proceedings of the Association for Information Science and Technology*, 57(1): e313.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Frischlich, L. 2020. #Dark inspiration: Eudaimonic entertainment in extremist Instagram posts. *New Media & Society*, 1461444819899625.

Gallup. 2020. Willingness to Get COVID-19 Vaccine Ticks Up to 63% in U.S.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gentzkow, M. 2016. *Polarization in 2016*. Toulouse Network for Information Technology Whitepaper.

Giachanou, A.; and Crestani, F. 2016. Tracking sentiment by time series analysis. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1037–1040.

Gilardi, F.; Gessler, T.; Kubli, M.; and Müller, S. 2022a. Social media and political agenda setting. *Political Communication*, 39(1): 39–60.

Gilardi, F.; Gessler, T.; Kubli, M.; and Müller, S. 2022b. Social Media and Political Agenda Setting. *Political Communication*, 39(1): 39–60.

Ginossar, T.; Cruickshank, I. J.; Zheleva, E.; Sulskis, J.; and Berger-Wolf, T. 2022. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations. *Human vaccines & immunotherapeutics*, 18(1): 1–13.

Gollwitzer, A.; Martel, C.; Brady, W.; Pärnamets, P.; Freedman, I.; Knowles, E.; and Van Bavel, J. 2020. Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour*, 4: 1186–1197.

Green, J.; Edgerton, J.; Naftel, D.; Shoub, K.; and Cranmer, S. J. 2020. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science advances*, 6(28): eabc2717.

Grimmer, J.; and Stewart, B. M. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267–297.

Hart, P. S.; Chinn, S.; and Soroka, S. 2020. Politicization and Polarization in COVID-19 News Coverage. *Science Communication*, 42(5): 679–697.

Hartman, R.; Blakey, W.; Womick, J.; Bail, C.; Finkel, E. J.; Han, H.; Sarrouf, J.; Schroeder, J.; Sheeran, P.; Van Bavel, J. J.; Willer, R.; and Gray, K. 2022. Interventions to reduce partisan animosity. *Nature Human Behaviour*, 6(99): 1194–1205.

Haworth, E.; Grover, T.; Langston, J.; Patel, A.; West, J.; and Williams, A. C. 2021. Classifying Reasonability in Retellings of Personal Events Shared on Social Media: A Preliminary Case Study with/r/AmITheAsshole. In *ICWSM*, 1075–1079.

Hetherington, M. J.; and Rudolph, T. J. 2015. *Why Washington won't work: Polarization, political trust, and the governing crisis*. Chicago, Il: University of Chicago Press.

Hong, S.; and Kim, S. H. 2016. Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, 33(4): 777–782.

Hu, T.; Wang, S.; Luo, W.; Zhang, M.; Huang, X.; Yan, Y.; Liu, R.; Ly, K.; Kacker, V.; She, B.; and Li, Z. 2021. Revealing Public Opinion Towards COVID-19 Vaccines With Twitter Data in the United States: Spatiotemporal Perspective. *Journal of Medical Internet Research*, 23(9): e30854.

Huddy, L.; and Yair, O. 2021. Reducing Affective Polarization: Warm Group Relations or Policy Compromise? *Political Psychology*, 42(2): 291–309.

Ikeda, K.; Hattori, G.; Ono, C.; Asoh, H.; and Higashino, T. 2013. Early detection method of service quality reduction based on linguistic and time series analysis of twitter. In *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, 825–830. IEEE.

Iyengar, S.; Lelkes, Y.; Levendusky, M.; Malhotra, N.; and Westwood, S. J. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22: 129–146.

Iyengar, S.; Sood, G.; and Lelkes, Y. 2012. Affect, not ideology: a social identity perspective on polarization. *Public Opinion Quarterly*, 76(3): 405–431.

Jang, H.; Rempel, E.; Roth, D.; Carenini, G.; and Janjua, N. Z. 2021. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. *Journal of Medical Internet Research*, 23(2): e25431.

Jones, R. H. 2011. Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, 30(25): 3050–3056.

Kantor, D. B. 2020. "A palpable fragmentation of Republican elites".

Larsson, A. O. 2020. Picture-perfect populism: Tracing the rise of European populist parties on Facebook. *New Media & Society*, 1461444820963777.

Lelkes, Y. 2016. Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80(S1): 392–410.

Liu, Y.-Y.; Tseng, F.-M.; and Tseng, Y.-H. 2018. Big Data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model. *Technological Forecasting and Social Change*, 130: 123–134.

López-de Lacalle, J. 2019. *R Package tsoutliers*, 1–2.

Marchal, N. 2021. "Be Nice or Leave Me Alone": An Intergroup Perspective on Affective Polarization in Online Political Discussions. *Communication Research*, 00936502211042516.

Matalon, Y.; Magdaci, O.; Almozlino, A.; and Yamin, D. 2021. Using sentiment analysis to predict opinion inversion in Tweets of political communication. *Scientific reports*, 11(1): 1–9.

Matamoros-Fernández, A.; and Farkas, J. 2021. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, 22(2): 205–224.

Mentzer, K.; Fallon, K.; Prichard, J.; and Yates, D. 2020. Measuring and Unpacking Affective Polarization on Twitter: The Role of Party and Gender in the 2018 Senate Races. *Hawaii International Conference on System Sciences 2020 (HICSS-53)*.

Moore-Berg, S. L.; Hameiri, B.; and Bruneau, E. 2020. The prime psychological suspects of toxic political polarization. *Current Opinion in Behavioral Sciences*, 34: 199–204.

Myers, S. A.; and Leskovec, J. 2014. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web*, 913–924.

Nanath, K.; and Joy, G. 2023. Leveraging Twitter data to analyze the virality of Covid-19 tweets: a text mining approach. *Behaviour & Information Technology*, 42(2): 196–214.

Noureen, S.; Atique, S.; Roy, V.; and Bayne, S. 2019. Analysis and application of seasonal ARIMA model in Energy Demand Forecasting: A case study of small scale agricultural load. In *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, 521–524.

Oliphant, J. B. 2020. Democrats see Biden and Sanders as very different ideologically.

Overgaard, C. S. B. 2021. Constructive Journalism in the Face of a Crisis: The Effects of Social Media News Updates About COVID-19. *Journalism Studies*, 22(14): Advance online publication.

Overgaard, C. S. B.; Masullo, G. M.; Duchovnay, M.; and Moore, C. 2022. Theorizing Connective Democracy: A New Way to Bridge Political Divides. *Mass Communication and Society*, 25(6): 861–885.

Pack, D. J. 1990. In defense of ARIMA modeling. *International Journal of Forecasting*, 6(2): 211–218.

Pew Research Center. 2020. Public opinion about coronavirus is more politically divided in U.S. than in other advanced economies.

Pierri, F.; Artoni, A.; and Ceri, S. 2020. Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections. *PloS one*, 15(1): e0227821.

Rathje, S.; Bavel, J. J. V.; and Linden, S. v. d. 2021. Outgroup animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26).

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851–860.

Samaras, L.; García-Barriocanal, E.; and Sicilia, M.-A. 2020. Comparing Social media and Google to detect and predict severe epidemics. *Scientific reports*, 10(1): 1–11.

Scheffer, M.; Borsboom, D.; Nieuwenhuis, S.; and Westley, F. 2022. Belief traps: Tackling the inertia of harmful beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, 119(32): e2203149119.

Settle, J. 2018. *Frenemies: How social media polarizes America*. New York, NY: Cambridge University Press.

Simchon, A.; Brady, W. J.; and Van Bavel, J. J. 2020. Troll and Divide: The Language of Online Polarization. *Working paper*.

Sims, C. A. 1986. Are forecasting models usable for policy analysis? *Quarterly Review*, 10(Win): 2–16.

Soroka, S.; and McAdams, S. 2015. News, politics, and negativity. *Political Communication*, 32(1): 1–22.

Team, C. 2020. CrowdTangle. Facebook, Menlo Park, California, United States.

Toda, H. Y.; and Phillips, P. C. B. 1994. Vector autoregression and causality: a theoretical overview and simulation study. *Econometric Reviews*, 13(2): 259–285.

Törnberg, P.; Andersson, C.; Lindgren, K.; and Banisch, S. 2021. Modeling the emergence of affective polarization in the social media society. *Plos one*, 16(10): e0258259.

Trussler, M.; and Soroka, S. 2014. Consumer Demand for Cynical and Negative News Frames. *The International Journal of Press/Politics*, 19(3): 360–379.

Tsao, S.-F.; Chen, H.; Tisseverasinghe, T.; Yang, Y.; Li, L.; and Butt, Z. A. 2021. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*, 3(3): e175–e194.

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political Polarization, and Political Disinformation: A Review of the Scientific Literature*.

Uyheng, J.; and Carley, K. M. 2020. Bot Impacts on Public Sentiment and Community Structures: Comparative Analysis of Three Elections in the Asia-Pacific. In Thomson, R.; Bisgin, H.; Dancy, C.; Hyder, A.; and Hussain, M., eds., *Social, Cultural, and Behavioral Modeling*, Lecture Notes in Computer Science, 12–22. Cham: Springer International Publishing. ISBN 978-3-030-61255-9.

van Atteveldt, W.; and Peng, T.-Q. 2018. When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, 12(2–3): 81–92.

van Atteveldt, W.; van der Velden, M. A. C. G.; and Boukes, M. 2021. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2): 121–140.

Wells, C.; Shah, D. V.; Pevehouse, J. C.; Foley, J.; Lukito, J.; Pelled, A.; and Yang, J. 2019. The Temporal Turn in Communication Research: Time Series Analyses Using Computational Approaches. *International Journal of Communication (19328036)*, 13: 1–22.

Westwood, S.; Iyengar, S.; Walgrave, S.; Leonisio, R.; Miller, L.; and Strijbis, O. 2018. The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research*, 57(2): 333–354.

Wilson, A. E.; Parker, V.; and Feinberg, M. 2020. Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences*, 34: 223–228.

Yarchi, M.; Baden, C.; and Kligler-Vilenchik, N. 2021. Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, 38(1–2): 98–139.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A as, we are not seeking to make population-level generalizations.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in the Ethics Statement.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in the Ethics Statement.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A as our study is exploratory and does not involve hypothesis testing.
   (b) Have you provided justifications for all theoretical results? Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No. The outliers identified are clearly identifiable to real-world events. No causal relationship is assessed in this study.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes.
   (f) Have you related your theoretical results to the existing literature in social science? Yes.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the code and data used to build the classifier is available via a URL. Due to the terms of services of the platforms, we do not provide a link to the full dataset used for the analyses. These datasets are available upon request.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Given our focus on classifying affective polarization in publicly available social media texts, we did not find reason to think that misclassification or intolerance would be harmful to any individuals.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? Yes, we've made our code and data publicly available and free to use under the CC-BY license, which is specified in the manuscript.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA. (We conducted aggregate analysis on publicly available data.)
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA. (We conducted aggregate analysis on publicly available data.)
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes. We're sharing our dataset and code through https://osf.io/. We have taken steps to ensure that our materials are freely available and as easy as possible to find and use for anyone.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Given the simplicity of the dataset we're releasing (columns: text and label), we did not find this to be neccesary.
6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.

(d) Did you discuss how data is stored, shared, and deidentified? NA.