Ebbs and Flows of Polarization During a Political Campaign

Kellin Pelrine[1,2], Anne Imouza[1,3], Gabrielle Desrosiers-Brisebois[1,3], Zachary Yang[1,2], Sacha Lévy[1,2], Aarash Feizi[1,2], Jiewen Liu[1,2], André Blais[3], Jean-François Godbout[3], and Reihaneh Rabbany[1,2]

[1]Mila
{kellin.pelrine, reihaneh.rabbany}@mila.quebec
[2]School of Computer Science, McGill University
{zachary.yang, sacha.levy, aarash.feizi, jiewen.liu}@mail.mcgill.ca
[3]Département de science politique, Université de Montréal
{anne.imouza, gabrielle.desrosiers-brisebois, andré.blais, jean-francois.godbout}@umontreal.ca

Abstract

This paper uses a joint embedding model, based on graph convolutional networks and RoBERTa, for textual and social network analysis, to estimate the partisan orientation of users of the micro-blogging sites Twitter and Parler during the US 2020 presidential election and its aftermath. By combining information on the structure of the social networks, likes, hashtags, re-posts, and the content of messages, the model estimates the partisan orientation of users who were active during the campaign. These estimates are based on over 2 million posts and user interactions taking place within them. We validate our results by using several gold standard measures, such as the voting records of Member of Congress and the party affiliation of users. We also present a novel approach for estimating the degree of partisan polarization on a daily basis throughout the campaign by grouping users along a left/right ideological continuum to analyze changes in cluster distances. Preliminary findings indicate that polarization increased after the 2020 election, with important shifts before the Capitol Hill riot.

Acknowledgement

Paper presented at the *Canadian Political Science Association* Meeting, June 9, 2021. This project is supported by CIFAR through a CIFAR AI Catalyst Grant: Being Politic Smart in the Age of Misinformation.

1 Introduction

The US 2020 election has been one of the most divisive in recent American history. The party system has become extremely polarized over the last thirty years, but this has reached unprecedented heights as Democrats and Republicans now strongly mistrust each other [Gelman et al., 2008, Iyengar et al., 2012, McCarty et al., 2016]. Our goal in this study is to document the ebb and flow of this partisan polarization by analyzing social media activities during and after the 2020 presidential election campaign.

There is a common understanding that social media platforms played an important role in increasing polarization around the election. With much of the campaign activities moving online because of the pandemic, these platforms provide rich data sources into the pulse of society, the public and elite alike. Motivated by this and inspired by the prior works of Barberá et al. [2015], Barberá [2015], Rheault and Cochrane [2020], and others, we examine the activities of users including the presidential candidates, Members of Congress and their followers, on the micro-blogging site Twitter, as well as the more conservative social media platform Parler.

More specifically, we collected around 350 million tweets and 6.5 million Parler posts, which reflects the activity of the mass public and the politicians (elite). Our data include xx individuals on Twitter, xx on Parler, as well as 2000 members of the elite on Twitter and 700 on Parler. Our data spans over several months, and we focus period just before

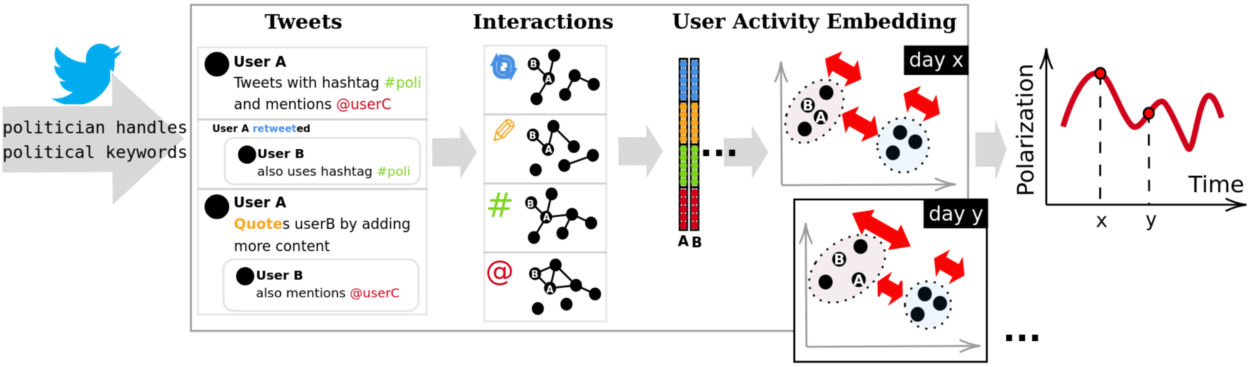

Figure 1: Mapping Users' Activity on Twitter to Create a Polarization Index

the November 3rd election to the end of the year. We record with whom social media users interact: quotes, re-posts and mentions. We also collect their hashtags usage as a proxy for their word choices as well as the actual text of their messages to fully monitor the language they used. We distinguish and study partisanship at the elite and at the mass levels to verify if our method works on both and we examine the differences between them. Figure 1 provides an overview of our methodology (explained in Section 3) to map the user's activity to polarization based on posts streamed from Twitter. The same method is applied to posts from Parler, although the terminology is different, e.g. retweet v.s. echo.

Our first task is to develop a measure of party polarization. We start at the elite level on Twitter. The set of political actors is defined as the Members of Congress and the two presidential and vice presidential candidates. We examine their 'conversations,' that is the words or hashtags they use on social media, as well as whom they refer to. From these, we estimate the location of each actor on an underlying dimension representing partisan conflict. Using a joint embedding model, we predict the party affiliation of politicians with very good accuracy. Our embeddings also contain information on the polarization of their voting records in Congress (as measured by DW-NOMINATE scores). These results indicate our model provides a valid representation off elite polarization.

We follow a similar approach at the mass level. We retrieve the Twitter posts of users related to the 2020 US election. We further identify Republicans and Democrats users from their profile by using a keyword filter and a language model classifier to construct a weak label of party identification. We validate this model on users labeled by experts, and then use its predictions to train a second model that learns from users' posts, rather than their profile, by following the same procedure as for the politicians. We show that this model is accurate at predicting the partisan orientation of non-elite users. This confirms that our model provides a reliable measure of mass polarization.

Our second task is to describe how party polarization fluctuates over time and to determine if events surrounding the election contributed to increase partisan polarization. Here, we are interested in identifying specific campaign and post-election moments that may have increased or reduced polarization. We also do various ablations, showing the impact of the different components of our model and the timescale.

The paper offers three main contributions. First, we present a new method to predict the party affiliation of social media users using a joint embedding model. Second, we propose a method to measure party polarization over time and at scale. And finally, using the above, we analyze changes in polarization around key events related to the 2020 US election and its aftermath, such as the many court decisions, or President Trump's decision to pardon Paul Manafort, Roger Stones, and Charles Kushner, while threatening to veto a major bipartisan Coronavirus relief bill.

Our paper is organized as follows. In the first section, we present a definition of partisan polarization on social media and review related work. The second section introduces the data and method used in the three main experiments of our analysis. The third section presents the results at the elite and mass levels, but also over time in a dynamic model after the 2020 presidential election. In the last sections we discuss these results and conclude.

2 Background

Partisan polarization has traditionally been measured by either looking at the difference between the policy positions of party members (spatial polarization) or by focusing on how much they dislike the other party (affective polarization). In spatial terms, polarization implies a movement away from the center towards the extremes, where distances are measured on a scale representing the left and the right of the ideological spectrum [Fiorina and Abrams, 2008]. In affective terms, partisan polarization focuses on citizens' emotions; it corresponds to the intensity of negative/positive feelings towards political parties [Gidron et al., 2020].

By using both measures, scholars have found that partisan polarization has increased over the last thirty years in the US, both at the elite level and in the mass public [e.g., Iyengar et al., 2012, McCarty et al., 2016, Gelman et al., 2008]. Elite polarization has primarily been confirmed over time by looking at trends in the behavior of Members of Congress through the spatial analysis of legislative voting records. Different scaling techniques of roll-call votes, like DW-NOMINATE scores [Poole and Rosenthal, 2007], have been used to estimate the ideological locations of Democrat and Republican representatives in a multidimensional policy space. Poole and Rosenthal [2007] have demonstrated that Representatives and Senators shifted their position to become increasingly distant from one another in recent years, which they take to be a sign of growing partisan polarization. Mass polarization, on the other hand, has been shown to exist in both affective and spatial terms through the analysis of public opinion survey data, either by comparing the policy preferences of Democrat and Republican party identifiers [Layman et al., 2006, Fiorina and Abrams, 2008] or by contrasting citizens' affective ratings of the different parties [Iyengar and Krupenkin, 2018, Hetherington and Rudolph, 2015].

Whether one looks at polarization in spatial or affective terms, measuring partisan divisions at the elite or mass level can either be done cross-sectionally, by looking at a specific point in time, or through time, by using multiple data points or panel data. As DiMaggio et al. [1996] and Fiorina and Abrams [2008] argue, it is particularly important to detect changes in polarization over time, as opposed to taking a snapshot of the distribution of policy preferences in a survey at one specific point in time, since a dynamic measure of partisan divisions can help us understand what type of events produces changes in public opinion.

The structure of data used in this analysis allows us to capture both the level and the change in partisan polarization before, during, and after the 2020 presidential campaign. By using social media information from the micro-blogging sites Twitter and Parler, we measure partisan polarization at the elite and mass public levels by identifying all of the Members of Congress and presidential candidates, and by looking at the behavior of users who are interested in US politics, which we assume to be a subset of partisans in the American population.

Our definition of partisan polarization combines both spatial and affective dimensions. We assume that partisanship is encoded in the choice of whom to follow/re-post/mention and similar social media interactions. These patterns of behavior relate to "homophilic" conversations between like minded individuals, as opposed to "heterophilic" conversations, which refer to the occasional exchanges between people who have weaker social ties [Yarchi et al., 2020, Barberá, 2015]. Our approach assumes that partisanship is related to word choices, that is which hashtags are used, which topics are discussed, and the vocabulary included in social media messages. The assumptions here are similar to the ones related to social interactions: like minded individuals tend to use the same type of vocabulary and the differences in word choices between Republicans (conservatives) and Democrats (liberals) should reflect partisan (ideological) divisions [Slapin and Proksch, 2008, Diermeier et al., 2012, Gentzkow et al., 2019]. Like Rheault and Cochrane [2020], our analysis is based on a neural network framework to combine both the information about the word choice and the network structure into a single model to estimate the ideological placement of social media users. We thus define:

> **Interactive polarization**, as the difference between the overall vocabulary and interactions observed within or across partisan groups. That is, the more partisans differ in their social media interactions or word choices, the greater the polarization. Conversely, the more they share or use similar language, the weaker the polarization.

This definition is in line with other approaches to estimate partisanship from text contained in online messages [Green et al., 2020, Grinberg et al., 2019, Gruzd and Roy, 2014, Yarchi et al., 2020]. It also naturally applies between two (or more) groups; for example between Democrats and Republicans or liberals and conservatives. It can be considered within a group as well, to identify people who are more extreme compared to the rest of the group. Below, we describe in greater details how we estimate individual polarization scores by examining text but also network interactions in a joint embedding model. However, we first review in the next section related work which estimates partisan polarization from social media data.

2.1 Related Work

Scholars have adopted three broad classes of models to measure partisan ideology from social network content. The first is based on the words used by users in their posts on social media. Here, researchers usually rely on a set of specific keywords in dictionaries to identify political messages and code their political leanings [Gruzd and Roy, 2014, Grinberg et al., 2019]. The political leanings can also be inferred from the text used in social media posts by using word embeddings [Conover et al., 2011, Yang et al., 2017]. This type of analysis is useful for detecting the main issues raised on Twitter (what people talk about), as well as the degree of partisanship contained in these messages (how they talk about it) [Green et al., 2020].[1] The second approach relies on the information provided by the network of users, who they follow, and who follow them in return [Conover et al., 2011]. This method is by far the most popular to infer the ideological leanings of social network users. Barberá [2015] offers the best examples of this type of analysis by estimating the left and right position of Twitter users through an item-response model, where the decision to follow a particular user is a function of ideology, the popularity of an account, and political interest. This model is then able to locate relevant ideological clusters on Twitter and confirms that users are more likely to interact with liked minded individuals.[2] Finally, a third group of models focuses more on the dynamic aspects of polarization in social networks over time by using either one of the two approaches described above. For example, Barberá et al. [2015] construct a daily index of polarization by relying on the network of users to demonstrate that certain events, like the Newtown Shooting in 2012, increased ideological conflict between liberals and conservatives on Twitter. On the other hand, authors like Green et al. [2020] use the text features of social media messages to build a dynamic measure of polarization over time. In this study, the authors trained a random forest machine learning algorithm to measure the level of elite polarization on Twitter during the 116^{th} Congress. Their results confirm that there was a surge in the level of polarization on COVID-19 related tweets, with Republicans becoming more distinctive in their behavior than Democrats in the early months of 2020.[3]

[1] Several studies [Gruzd and Roy, 2014, Yang et al., 2020, Grinberg et al., 2019, Yang et al., 2017] have attempted to measure the ideological orientation of Twitter users by looking at the specific textual content of their messages. Some of these studies [Gruzd and Roy, 2014, Grinberg et al., 2019] have relied on sets of specific keywords to infer political tweets and their sentiment/political leaning. Similarly, Yang et al. [2017] and Yang et al. [2020] have relied on semantic representation of hashtags using the "word2vec embedding" in order to measure the average difference between or within specific tweets aggregated by groups to infer users' ideological alignment on a left-right scale. Moreover, another study [Badawy et al., 2018] has determined the political ideology of Twitter users based on the political leaning of the media outlets they shared on their profiles.

[2] Another closely related approach relies on defining general ideal points of moderate Democrat and Republican Senators by using roll call data [Chen, 2015]. Other studies have looked at the ideological distance between users by observing patterns of interaction among party followers in Europe [Bright, 2017, Gaisbauer et al., 2021].

[3] Other studies like Yardi and Boyd [2010] study the issues of gun violence and abortion on Twitter over a period two months and state that homophily may impact polarized discussions on-

In this study, we propose a new method that combines all three of these approaches into a single model using a joint embedding framework. Our goal is to estimate the ideology of social media users by looking at the content of their messages and their networks, but also to determine if party polarization has fluctuated over time during the most recent presidential election.

3 Method

In this section, we describe how we collected the data on the social networks Twitter and Parler. We also explain how we identified elite and non-elite users to construct measures of partisanship. Finally, we discuss the methodology for each estimating the partisan orientation of each user, which serve as a basis to develop our dynamic measure of polarization during and after the 2020 election campaign.

3.1 Data Collection

We curated four datasets, summarized in Table 1. In this table, the users are the authors of the posts, while the nodes represent users in our interaction graphs (see Section 3.2.1)—the authors plus users that are referenced within the posts. The hashtag, mention, retweet (or re-post), and quote columns, indicate the number of edges connecting the nodes.

Twitter We collected all tweets, retweets and replies from 995 elite accounts linked to the public and personal Twitter accounts of the US representatives (433), senators (99), as well as vice presidential and presidential candidates (8) using Twitter's Search API.[4] We call this the Politicians dataset.

We also collected around 1% of real-time tweets using Twitter's streaming API, that included one of the following US election related keywords: [JoeBiden, DonaldTrump, Biden, Trump, vote, election, 2020Elections, Elections2020, PresidentElectJoe, MAGA, BidenHaris2020, Election2020]. This constitutes the Election dataset with approximately 350 million tweets and 20 million users. From these, we sampled 20 thousand users, which is the mass Public dataset.

Some days in the Public dataset are missing due to interruptions in the collection pipeline: October 28th, November 17th and 24th, and December 1st, 12th, 13th, 22nd, and 23rd. These days are not shown in our results. There are also two days that are partially missing, December 2nd and 9th. These days have about half the expected amount of data, which is sufficient for meaningful measurements, so we include them.

Parler We parsed all posts provided by the Distributed Denial of Secrets[5] and WayBack Machine[6]. Posts parsed[7] have an estimated creation date since the data provided contain relative timestamps such as “1 day ago” or “1 week ago”. Parler posts (or Parleys) can contain hashtags (#) and re-posted content (echo).

DW-NOMINATE In order to get an exogenous measure of partisan ideology for our elite group of users, we use Poole and Rosenthal [2007]'s DW-NOMINATE scores for House Representatives and Senators who served in the 116^{th} Congress.[8] These scores are obtained from the roll call votes of Members of Congress through a multidimensional scaling procedure. The projected first dimension has been shown to represent the ideological conflict opposing the left and the right—from the most extreme to the most moderate positions. Each member is aligned on this continuum, depending on how liberal or conservative their voting record is. These scores were then matched to the Politicians profiles present on Twitter and Parler.

line. Badawy et al. [2018] estimate a dynamic measure by asserting a political leaning to each user regarding the media outlets they share over a period of two months before the 2016 US Presidential Election. Garimella and Weber [2017] also investigate changes in political polarization on Twitter between 2009 and 2016 by estimating the ideology of users from the type of politicians and media they follow. Their results confirm that polarization has increased over time.

[4] Some Members of Congress have more than one social medial account (e.g., one personal and one official account). In this case, we collected information for all of the relevant accounts.

[5] https://ddosecrets.com/wiki/Parler

[6] https://web.archive.org/web/*/https://parler.com

[7] https://github.com/RSTZZZ/parler_parser

[8] https://voteview.com

Dataset	Posts Collected					Interaction Graphs				
	Source	Start	End	Posts	Users	Nodes	Hashtag	Mention	Retweet	Quote
Politicians	Twitter	2020-08-01	2021-01-17	156,562	995	19,423	1,124,621	2,435,055	134,333	269,104
Public	Twitter	2020-10-26	2020-12-31	2,871,050	20,008	244,931	1,054,431	11,150,470	2,016,192	1,010,138
Election	Twitter	2020-10-26	2021-01-04	348,671,076	20,533,417	-	-	-	-	-
Parler	Parler	2020-10-25	2021-01-08	6,546,658	566,486	-	-	-	-	-

Table 1: Statistics on the collected datasets and the interaction graphs extracted from this data. For example, we have 4 graphs of 19,423 nodes representing the interactions that our 995 politicians had with other users in Twitter.

3.1.1 Classification

In order to train and evaluate our models, we classify a sample of users according to their party affiliation and ideology based on the description they provide on their user profile on Twitter and Parler.

First, for each dataset, we classify users as “conservative”, “liberal” or “unknown” based on identifiers in the description. For “conservative,” we use: [conservative, gop, republican, trump]. For “liberal,” we use: [liberal, progressive, democrat, biden]. We label users as “conservative” (“liberal”) if the description contains at least one of the conservative (liberal) identifiers and does not include any of the liberal (conservative) identifiers. The rest of the users remain as “unknown.”

This is a “weak” classification because user keywords may not match their actual party affiliation or ideology. For example, instead of a president name indicating support, they could say ”I hate Trump” or ”I hate Biden.” In order to validate the overall performance of these labels, we asked two expert coders to classify, on the basis of the very same information (that is, the description provided in the user profile) 60 users from the politicians dataset, and 1000 general public Twitter users from each party, 200 conservative Parler users, and 500 liberal Parler users. This “strong” classification either confirms the weak labels, or indicates the presence of a coding error. Note that while in most cases an incorrect weak liberal label indicates that the user is in fact a conservative (or vice versa), a small number of these users can also be independent or apolitical. After comparing the weak with the “strong” labels, we found that users in the Politicians dataset are generally more politically involved and hence the simple keyword search is very accurate. However, for other users, the accuracy was lower, with only around 70% of the weak labels matching the strong labels.

Therefore, we used the strong labels to train our classifier to generate a more accurate classification. We randomly split the strong-labeled data into a 75% training set with Twitter and Parler combined, and a separate 25% test set for each platform. With this data we fine-tuned a roberta-large [Liu et al., 2019] model to predict the party each user is closest to from their profile description.[9] We report the results in Table 2.

Dataset	Counts		Accuracy	
	Cons.	Lib.	Cons.	Lib.
Politicians	1,174	1,068	97.7%	96.8%
Election	183,207	176,271	87.0%	90.5%
Parler	31,966	808	93.1%	82.9%

Table 2: User self-declared party/ideological alignment and accuracy of our classifier

These results show that the classifier provides a reasonably accurate classification of party labels. They are still imperfect, but they are sufficiently accurate for use in training our model. We note that the classifier is binary, liberal or conservative—it cannot classify a user as moderate or independent. In situations where one of those labels would be more appropriate, the classifier will nonetheless say whether it thinks the user is closer to being a liberal or a conservative. We also note that there are far more conservative than liberal users on Parler, matching the platform’s reputation for being almost exclusively favored by conservatives.

[9] RoBERTa is a pretrained language model; the large version we use has 355 million parameters. It is based on the transformer [Vaswani et al., 2017] and the BERT architecture [Devlin et al., 2018], with modifications designed to improve the training process.

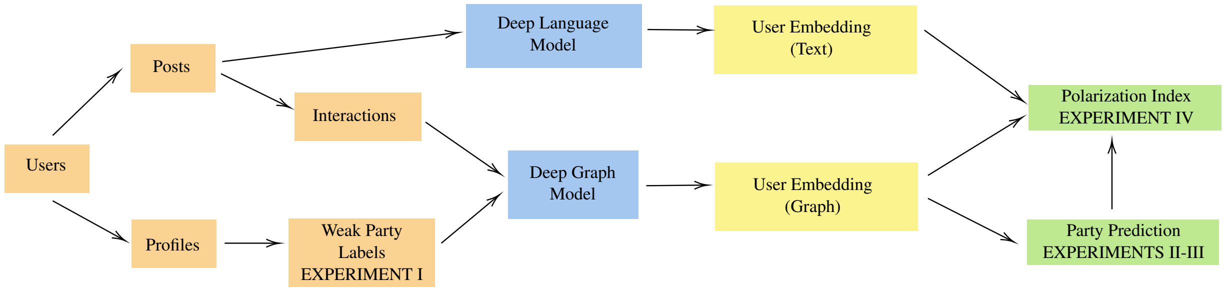

Figure 2: Leveraging users with explicit profile information for mass ideology prediction.

3.2 Measuring User Activity

An overview of our approach is shown in Figure 2. We collect each user's posts and profile descriptions. From their profiles, we construct weak labels. From their posts, we extract user interactions. These generate a series of graphs which are then integrated into different deep graph models to produce user activity embeddings. Meanwhile, we also integrate the posts into a deep language model to get user activity embeddings based on the text. We use these embeddings to predict party affiliation and measure partisan polarization.

3.2.1 Constructing Interaction Graphs

The next step in assembling our data is to construct interaction graphs—i.e., graphs where the nodes are users and edges represent interactions between them. There are two types of interactions that can link people together:

- Direct links. For example, person A mentions or re-post person B.
- Indirect links. For example, persons A and B both mention the same person C, even though A and B may not mention each other directly.

Direct links are intuitive, however, we find empirically that indirect links are more useful for grouping similar and separating dissimilar partisans. Therefore, we construct graphs with edges based on indirect links between user nodes.

We construct one graph each for hashtags, mentions, retweets, and quotes on Twitter (note: the Parler analysis will be included in a subsequent version of this paper). These interactions can be collected directly from tweet data from the Twitter API. Previous works such as Barberá [2015] heavily used follower networks, but these require separate scraping that can be challenging on a large scale.

In order to get accurate predictions and measurements, some amount of user activity is needed—if a user is not connected to anyone else in the network, our model cannot give a meaningful prediction. Therefore, for a user to enter our interaction graphs we require at least 10 edges connecting them to other users (for example, they use a hashtag which is also used by 10 other people). Second, we filter users who appear in all four interaction graphs. This filter is applied exclusively on the output side—the other users still appear in the training set, but they are not used to evaluate the results.

Most of the politicians are quite active, so we retain 724 accounts of members of congress. There is much more variation in the mass public, but we retain approximately between 100-250 strong-labeled users every day, which we use for evaluating party prediction. When measuring polarization over time, we do not require strong labels, which gives approximately 500-1000 users per day.

3.2.2 Generating User Embeddings

We show the modeling process in Figure 3.

We first estimate an ideological position with the text of each user's posts. We use a roberta-large language model (LM). In this case, we want an embedding for each user rather than an immediate prediction of their party, so we use the pretrained model directly, without fine-tuning on any prediction task. We embed all the tweets, then average the embeddings per user to get a single embedding for each

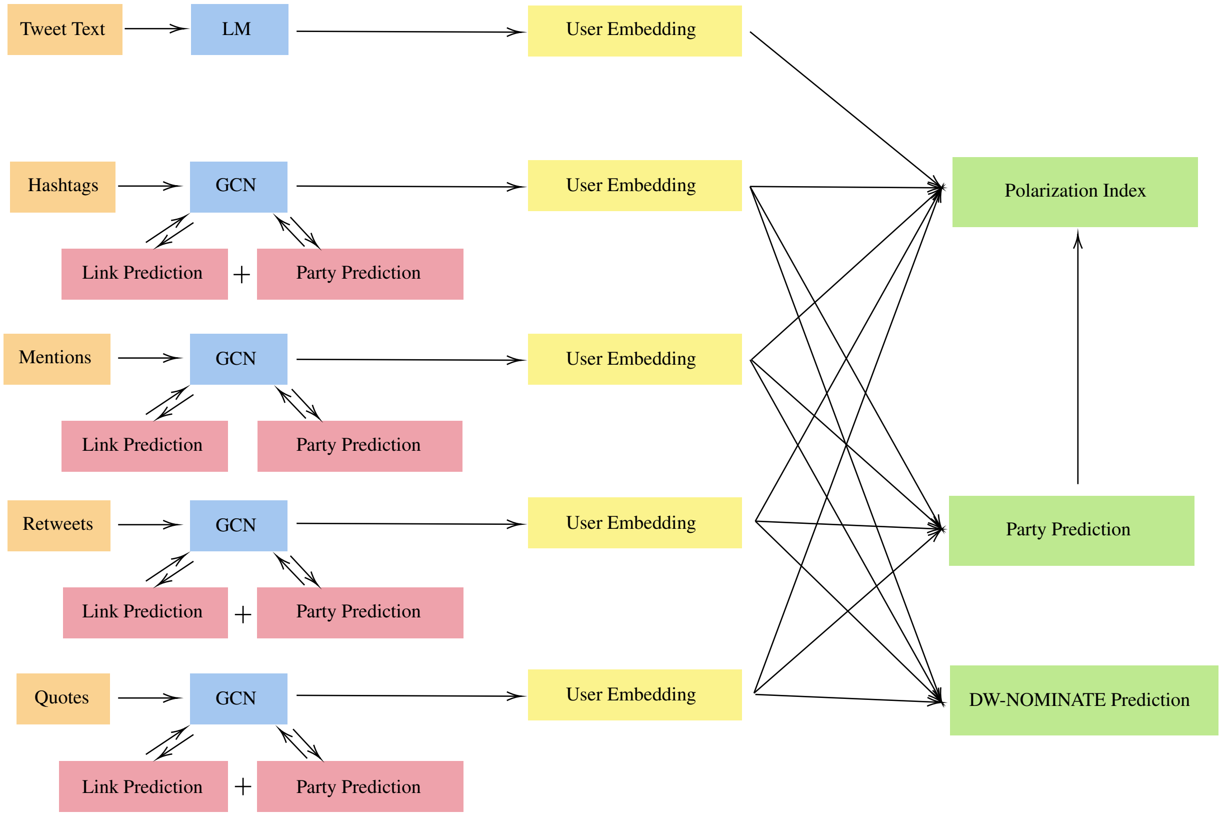

Figure 3: We use a multimodal method to combine 5 types of data into party predictions and polarization measurements

user.

Next, we use the interaction graphs. We do this with a semi-supervised graph convolutional network (GCN) [Kipf and Welling, 2017]. Our model has two terms in the loss function used to train the model, added together with equal weight: one for link prediction, and one for party prediction. For link prediction, we construct a randomly connected negative graph, as in a Deep Graph Library example.[10] The model learns from this by trying to predict which links are from the real graph and which are from the negative one, using a cross entropy loss.

With regards to the party prediction part, for the politicians, we use the weak labels where available for non-politician users, and the true labels for politician users in the training set (which essentially match the weak labels, since they are very accurate on this data, as shown in Table 2). For the public, we use the weak labels exclusively, where available. The model learns from them with another cross entropy loss.

From this model, we get one embedding per user and per interaction type. We can combine these in various ways to predict ideology, party, and measure polarization, as discussed in the following section.

GCNs typically use node features. We initially planned to use the content of the text as discussed above. However, we found empirically that it did not seem to provide any clear benefits. This matches the findings of Xiao et al. [2020]. We simply use a random vector as node features. However, while the text of the text does not appear to improve link or party prediction beyond the interaction graphs themselves, or at least not that the GCN is extracting effectively, it likely still contains useful information on user polarization. Therefore, rather than as node features, we use the text embeddings directly as a fifth type of representation.

We train our GCNs for 1000 epochs, chosen empirically with validation data. We use the Adam optimizer [Kingma and Ba, 2017]. All models are trained on an RTX8000 GPU. For each GCN,

[10] https://docs.dgl.ai/en/0.6.x/guide/training-link.html

we produce a 100-dimensional embedding, while roberta-large produces a 1024-dimensional embedding.

3.3 Measuring Ideology, Partisanship, and Polarization

3.3.1 Politician Party Polarization

To predict politician party, we concatenate the user activity embedding and pass them to a random forest classifier model, implemented with default settings through scikit-learn [Pedregosa et al., 2011]. We preserve the same train-test split as the GCN.

We predict DW-NOMINATE scores similarly, with a random forest regression model this time. We find empirically that it is important to include in this model the politicians' parties as an input. This parallels Barberá [2015], who gives party label in the initialization stage of his analysis.

We observed that AutoGluon [Klein et al., 2020] can often produce slightly stronger performance than random forests. However, it is much slower, which can be troublesome: in one of our analysis, we do 100 runs to reduce random variance in the final measurements. Therefore, we have decided to use the efficient random forest approach and we plan to investigate AutoGluon in future work.

3.3.2 General Public Party

For the general public, we follow the same modeling process as for predicting politician party. In order to measure polarization over time, besides the overall party prediction, we also produce predictions for each day individually. This is done by filtering our data down to a single day, running the pipeline on that specific time point and saving the results, then proceeding to the next day.

3.3.3 Public Polarization Over Time

Given the daily user activity embeddings and predicted party label for each user from the previous section, we can measure polarization using a cluster quality metric. Intuitively, the better separated the two clusters are (corresponding to the ideological division between liberals and conservatives), the more polarized these positions are.

For quality metric, we use C-index, which compares the dispersion of clusters of data relative to the total dispersion found in the dataset [Hubert and Levin, 1976]. C-index is one of the best performing criteria used for the validation of clustering results [Rabbany et al., 2012, Vendramin et al., 2010]. More formally, it is computed as:

$$C = \frac{S_{max} - S_w}{S_{max} - S_{min}} \tag{1}$$

Here, S_w is the sum of within-cluster Euclidean distance measurements, which we assume to be linked to spatial polarization [Poole and Rosenthal, 2007]. S_{min} is the sum of the smallest distances between points. S_{max} is the sum of the largest distances between points. A higher C-index corresponds to more dispersed data—i.e., higher polarization.

We first compute the C-index using each type of user activity embeddings individually. This gives five series of C-index, one for each of the interaction types (hashtag, mention, retweet, and quote), and one for the text. Note the relative scale of each is hard or impossible to interpret. For example, we see that the average number for mentions is high compared to text, but this does not necessarily mean mentions are more polarized than text, because they are computed with different models and from different types of data. Rather, it is changes over time, within each series, that are meaningful. If the C index for mentions increase over some days, then that indicates there is increasing polarization in who the users are talking to or the vocabulary they use.

We combine the individual indices into our single overall polarization index by taking the product. We choose this aggregation because it is simple and clear. It reflects the intuition that each component contains useful information, and that an aggregate metric that is proportional to the individual ones is reasonable. For example, if polarization in text increases while polarization in mentions decreases, both of these changes are important and should be accounted for in our measurement. With the product aggregation, one of the two changes will dominate if it is a proportionally bigger change, or they will cancel out if they are proportionally similar.[11]

[11] Two other simple options are taking the average or the max-

4 Findings

This section presents our main findings. We group the analysis in three parts:

> **Question 1:** are our user measures from social media activity **predictive of known politician's positions?** Politicians' ideological positions are available based on their voting records. Here we provide a (static) evaluation of our own measures for this specific set of users, politicians, for which we know their DW-NOMINATE scores. We show our embeddings are meaningful by predicting these scores, as well as the politicians' parties.

> **Question 2:** Are our measures **predictive of the ideology of the mass public?** We show our embeddings also for the mass public, using weak labels (i.e. labels that may be imperfect, for example from keywords like "liberal" or "conservative" in the user profile).

> **Question 3:** Does polarization **increase around major polarizing events?** We use our embeddings to examine changes over time. We show that changes in polarization correspond to real world events (such as the date of the election or major political events). We also use a synthetic model where we can control how much people are "talking to" each other to further understand and validate the results.

4.1 Results

4.1.1 Politicians

In Table 3 we predict NOMINATE scores and measure the correlation between our own measure and NOMINATE scores. We find that there is high variance depending on the train-test split and training of the random forest. Therefore, we report the average and standard deviation of 100 runs with 70-30 random splits. We have a high correlation overall and a weaker but positive correlation within each party.

All	Democrat Only	Republican Only
0.96 ± 0.00	0.27 ± 0.07	0.33 ± 0.08

Table 3: user activity embedding predicts ideological scores of members of congress. Table reports the Pearson correlation between DW-NOMINATE scores and our measure.

In Table 4 we show the accuracy of our method in predicting politician's party affiliation. First, we report accuracy using each interaction relation individually (Hashtag, Mention, Retweet, and Quote), and then the accuracy when combining all four. The latter is significantly higher.

Hashtag	Mention	Retweet	Quote	Combined
75.0	81.5	75.2	71.1	91.2

Table 4: We predict politician party with high accuracy (%), especially with the combined model user activity embedding

4.1.2 Mass Public

Table 5 is similar to Table 4, but with the Public dataset. It shows our method is accurate.

Hashtag	Mention	Retweet	Quote	Combined
85.3	88.0	92.6	82.9	93.1

Table 5: We predict general public party accurately (%)

We performed analyses comparing our model to the TIMME model of Xiao et al. [2020]. On November 3^{rd} alone, our model achieves 91.7% prediction

imum value. However, these can introduce scaling issues, because as noted in the preceding paragraph, the scale itself is hard to interpret between the different relations. For example, if one relation produces C-index which fluctuates within a small range, while another fluctuates in a large range, then the latter can dominate the former in the average. And similarly, if one relation produces C-index which is consistently higher than another, then the latter will be ignored in the maximum. In addition, the maximum can magnify noise in the measurements—the more relations one combines, the more likely the maximum will be just the one with the largest noise. By using the product, we avoid these potential pitfalls. In future work, we will consider more complex ways of combining the embeddings and distance measure.

accuracy, while TIMME achieves 93.1% accuracy. Thus, TIMME is slightly more accurate on that particular day, but our performance is still comparable. However, TIMME was much slower, taking approximately three hours for the single day compared to 30 minutes for our model, on exactly the same hardware. In addition, we have observed cases where our model can run on a larger amount of data, such as the full Public dataset used in Table 5, while TIMME runs out of memory, again with the same hardware. Thus, our accuracy is quite good and our model is significantly more scalable. This scalability is critical, considering the size of our data and our goal of measuring polarization, not just on one day, but over time. In future work, we will perfom additional analyses to compare the two models.

4.1.3 Polarization over time

In this final section, we consider results over time. First, in Figure 4, we present party prediction accuracy. This is similar to Table 5, but while those results were using the entire Public dataset, here we take each day and make a prediction on that day's data alone, with the combined model. Our model achieves a consistently high accuracy.

The vertical lines, here and in the subsequent figures, represent November 3^{rd} (election day) and December 23^{rd} (the day electoral votes arrived in the capital, as well as other events discussed in the following section).

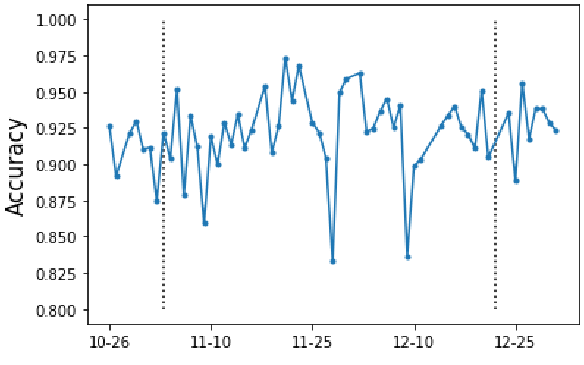

Figure 4: We predict party accurately even on individual days

In Figure 5, we show the C index over time for each type of user embedding. Each line can be regarded as a particular kind of polarization. For example, the text line measures the polarization of the language people use, while the mention line represents how polarized their social interactions are.

In Figure 6, we present our overall daily polarization index, which is an aggregation of the indices in the previous figure. This single index is easier to understand than the previous figure. We discuss how it evolves in the following section.

5 Discussions

The overall level of correlation observed between our measures and the NOMINATE scores, shown in Table 3, is very high (.96). While this seems excellent on the surface, we note that a similar correlation can be obtained from linear regression on the politician party alone. However, our method achieves intraparty correlations of .27 (Democratic) and .33 (Republican). This is despite using completely different data sources—either social media activities or NOMINATE scores, which are derived from roll-call votes. While these correlations are not perfect, they show our user activity embeddings are capturing partisan polarization.

We can also accurately predict politician's party, as seen in Table 4, again showing that our user activity embeddings are meaningful. We see a significant improvement from using all four interaction types together, motivating the use of a multimodal method like ours.

Cohen and Ruths [2013] found that a good performance for the political elite is insufficient to guarantee fo good performance for the general public. In Table 5, we see that our method still works effectively when we focus on the mass public. In addition, despite reducing the size of the data, Figure 4 confirms that this performance holds when we move to the temporal setting and apply the model on each day individually.

Finally, we find in Figure 6 that polarization is changing over time on Twitter. First, the results indicate that there was a slight decrease in partisan polarization observed in the days immediately following the November 3^{rd} election. However, we find that the overall level of polarization subsequently increased throughout the month, as President Trump

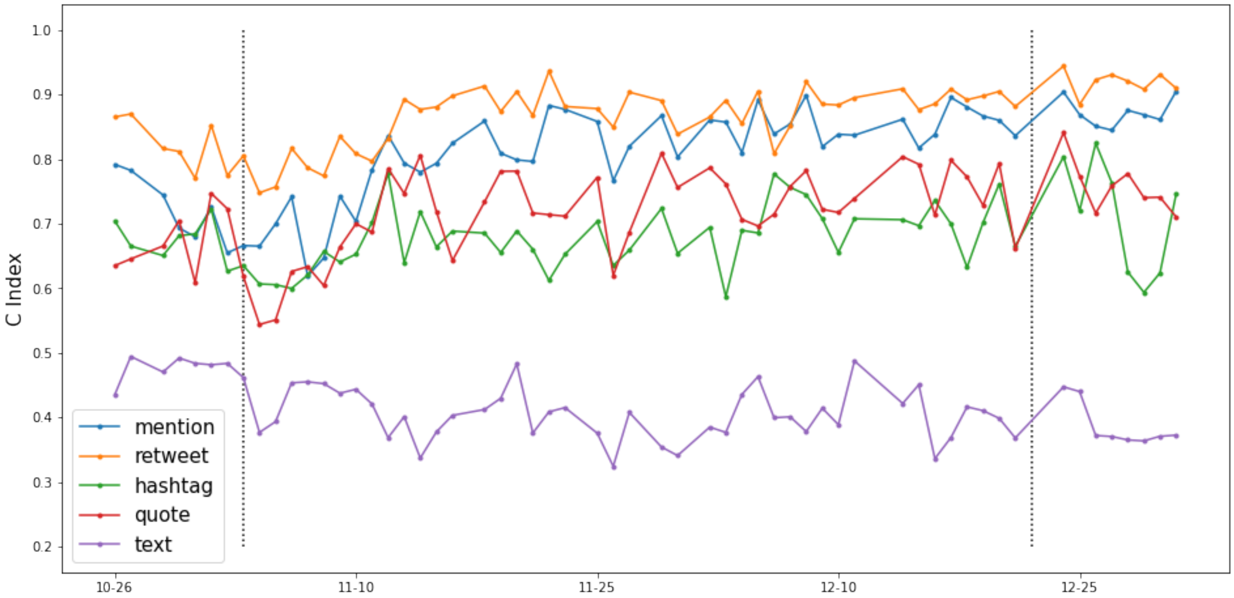

Figure 5: General public polarization over time, using each type of data individually

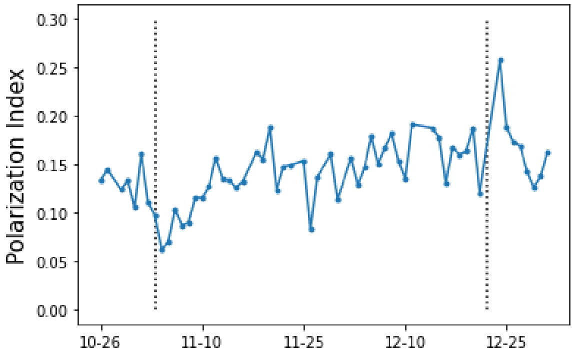

Figure 6: General public polarization over time. Polarization decreases around the election (1st vertical line), then increases, and peaks on December 24^{th}.

and his supporters continued to contest the legitimacy of the vote. The Figure also suggest that there was a significant surge in partisan polarization around December 23^{rd}. Several important political events occurred on this day, such as the presidential pardon of Roger Stone, Paul Manafort, and Charles Kushner. At the time, President Trump also attempted to block the adoption of two bipartisan pieces of legislation, the Defence spending bill and the COVID rescue package. It remains easy to find meaningful events that correspond to peaks in the polarization estimates found in Figure 6, especially since President Trump's behavior was so erratic in the weeks following the election. We must be careful in this type of post hoc "real world" event analysis. In future work, however, we plan to conduct additional tests to confirm the predictive validity of our model.

6 Conclusions and Future Works

This paper introduced a new dynamic framework to study political polarization on social media platforms. This approach is the first to combine both social interactions and the text content of online messages to estimate a measure of partisan polarization by analyzing more than 365 millions posts on Twitter and Parler during the 2020 presidential campaign. We do this by generating user activity embeddings with a deep language model (roberta-large) and deep graph models (GCNs). In our analyses, we showed that these user activity embeddings effectively capture information on ideology and polarization.

Applying them to measure general public polarization over time, our findings confirm that there was a small decline in partisan polarization after the November 3^{rd} election, followed by a gradual increase in partisan conflicts in the following weeks, with President Donald Trump and his supporters challenging the election results. Our findings also show that polarization increased significantly a little before Christmas Eve, when President Trump pardoned Roger Stone, Tom Manafort, and Charles Kushner, and threatened to veto a largely popular COVID stimulus bill. We expect this polarization to increase even more in January, especially around January 6^{th}, the day of the infamous Capitol Hill Riots. Unfortunately, at the time of writing this paper, our data collection task for this period was incomplete, so we leave this analysis for future work. Below are some of the additional tasks we plan to do in the next version of the paper:

- Run the party classifier and the geolocation classifier on the whole dataset, to get the proportion of Republicans and Democrats in different states and cross reference with statistics known for these states.
- Examine other samples of the general public from our large dataset, and construct bootstrap confidence intervals for the measurements.
- Construct a dynamic measure of polarization for Parler users.
- Classify social media users that are independent or non-partisan into a separate analytical category.
- Expand our analysis to include more days preceding the November 2020 election. Likewise, expand our analysis in the future to include more days in the first few months of 2021. Conduct additional predictive validity tests to check if meaningful events are linked to significant changes in polarization.

In the future, we also hope to improve the accuracy of our dynamic measure of political polarization and expand its usage to different social media networks, such as Facebook, Reddit, Instagram and TikTok. Our goal is to develop a model that is scalable to very large datasets and applicable to alternative social media platforms. Finally, we plan to use our measure of polarization in a comparative study to monitor upcoming elections across different countries, including Canada.

References

A. Badawy, E. Ferrara, and K. Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265. IEEE, 2018. URL `https://arxiv.org/pdf/1802.04291.pdf`. 4, 5

P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91, 2015. 1, 3, 4, 7, 9

P. Barberá, J. Jost, J. Nagler, J. Tucker, and R. Bonneau. Tweeting from left to right. *Psychological Science*, 26:1531 – 1542, 2015. 1, 4

J. Bright. Explaining the emergence of echo chambers on social media: the role of ideology and extremism, 2017. 4

Z. Chen. Mass ideology-based voting model. In *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 871–877, 2015. doi: 10.1109/IAEAC.2015.7428681. 4

R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In *Seventh international AAAI conference on weblogs and social media*, 2013. 11

M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE, 2011. 4

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`. 6

D. Diermeier, J.-F. Godbout, B. Yu, and S. Kaufmann. Language and ideology in congress. *British Journal of Political Science*, pages 31–55, 2012. 3

P. DiMaggio, J. Evans, and B. Bryson. Have american's social attitudes become more polarized? *American journal of Sociology*, 102(3):690–755, 1996. 3

M. P. Fiorina and S. J. Abrams. Political polarization in the american public. *Annu. Rev. Polit. Sci.*, 11: 563–588, 2008. 3

F. Gaisbauer, A. Pournaki, S. Banisch, and E. Olbrich. Ideological differences in engagement in public debate on twitter. *Plos one*, 16(3): e0249241, 2021. 4

V. Garimella and I. Weber. A long-term analysis of polarization on twitter. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 528–531. AAAI PRESS, 2017. URL `https://www.icwsm.org/2017/`. International AAAI Conference on Web and Social Media, ICWSM ; Conference date: 15-05-2017 Through 18-05-2017. 5

A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, et al. A weakly informative default prior distribution for logistic and other regression models. *Annals of applied Statistics*, 2(4):1360–1383, 2008. 1, 3

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. 3

N. Gidron, J. Adams, and W. Horne. *American Affective Polarization in Comparative Perspective*. Cambridge University Press, 2020. 3

J. Green, J. Edgerton, D. Naftel, K. Shoub, and S. J. Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. *Science Advances*, 6(28):eabc2717, 2020. 4

N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378, 2019. ISSN 0036-8075. doi: 10.1126/science.aau2706. URL `https://science.sciencemag.org/content/363/6425/374`. 4

A. Gruzd and J. Roy. Investigating political polarization on twitter: A canadian perspective. *Policy & Internet*, 6(1):28–45, 2014. doi: 10.1002/1944-2866.POI354. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI354`. 4

M. J. Hetherington and T. J. Rudolph. *Why Washington won't work: Polarization, political trust, and the governing crisis*, volume 104. University of Chicago Press, 2015. 3

L. Hubert and J. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1080, 1976. 9

S. Iyengar and M. Krupenkin. The strengthening of partisan affect. *Political Psychology*, 39:201–218, 2018. 3

S. Iyengar, G. Sood, and Y. Lelkes. Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431, 2012. 1, 3

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. 8

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, 2017. 8

A. Klein, L. Tiao, T. Lienart, C. Archambeau, and M. Seeger. Model-based asynchronous hyperparameter and neural architecture search. *arXiv preprint arXiv:2003.10865*, 2020. 9

G. C. Layman, T. M. Carsey, and J. M. Horowitz. Party polarization in american politics: Characteristics, causes, and consequences. *Annu. Rev. Polit. Sci.*, 9:83–110, 2006. 3

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. 6

N. McCarty, K. T. Poole, and H. Rosenthal. *Polarized America: The dance of ideology and unequal riches*. mit Press, 2016. 1, 3

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 9

K. T. Poole and H. Rosenthal. On party polarization in congress. *Daedalus*, 136(3):104–107, 2007. 3, 5, 9

R. Rabbany, M. Takaffoli, J. Fagnan, O. R. Zäane, and R. J. Campello. Relative validity criteria for community mining algorithms. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 258–265. IEEE, 2012. 9

L. Rheault and C. Cochrane. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1): 112–133, 2020. doi: 10.1017/pan.2019.26. 1, 3

J. B. Slapin and S.-O. Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008. 3

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6

L. Vendramin, R. J. Campello, and E. R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical analysis and data mining: the ASA data science journal*, 3(4):209–235, 2010. 9

Z. Xiao, W. Song, H. Xu, Z. Ren, and Y. Sun. Timme: Twitter ideology-detection via multi-task multi-relational embedding. *arXiv preprint arXiv:2006.01321*, 2020. 8, 10

K.-C. Yang, P.-M. Hui, and F. Menczer. How twitter data sampling biases u.s. voter behavior characterizations, 2020. 4

M. Yang, X. Wen, Y. Lin, and L. Deng. Quantifying content polarization on twitter. In *2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*, pages 299–308, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/CIC.2017.00047. URL `https://doi.ieeecomputersociety.org/10.1109/CIC.2017.00047`. 4

M. Yarchi, C. Baden, and N. Kligler-Vilenchik. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, pages 1–42, 2020. 3, 4

S. Yardi and D. Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of science, technology & society*, 30(5):316–327, 2010. 4