Online Appendix for "280 Characters of Contention: Analyzing Partisan Behavior on Twitter During Supreme Court Confirmation Processes"

Search Terms used for Downloading Tweets

	Kavanaugh	Barrett	Jackson
Vacancy	Justice Kennedy, Kennedy, #SCOTUS-nominee, #midtermsbeforescotus	RBG, Ginsburg, SCOTUS	Breyer, SCOTUS, Justice Breyer, Stephen Breyer
Nomination	Brett Kavanaugh, Trump AND Kavanaugh	Amy Barrett, Amy Coney Barrett, SCOTUS	KBJ, Ketanji, Brown AND Jackson, Judge AND Jackson, SCOTUS, Supreme Court
Hearings	#SCOTUS, Confirmation Hearing, Kavanaugh	Amy Barrett, Amy Comey Barrett, SCOTUS, Confirmation Hearing	Confirmation Hearing, KBJ, Ketanji, Judge AND Jackson, SCOTUS, Supreme Court, Supreme Court hearings
Assault Hearings	Supreme Court, Kavanaugh, Kavanaugh AND Ford, Christine Ford, #MeToo, Dr. Ford, Deborah Ramirez, #KavanaughHearings, #BelieveSurvivors		
Vote	Kavanaugh	Amy Barrett, Amy Coney Barrett, SCOTUS, Confirmation Vote	#KetanjiBrownJackson, Confirmation vote, Judge AND Jackson, KBJ, Ketanji, SCOTUS, Supreme Court

Table A1: Search terms used in automated TAGs software. Terms containing phrases were enclosed in quotation marks to ensure only phrase hits were returned.

Sentiment Analysis

Tweets Used for Sentiment Analysis

		Tot.	Vacancy	Announc-ement	Conf. Hearing	Ford Hearing	Final Vote
BMK	Dem	4242	612	634	1353	810	833
	Rep	1338	229	283	360	141	325
	-	159797	28631	27631	34443	35129	33923
	All	165337	29472	28548	36156	36080	35081
ACB	Dem	7073	768	1448	4130	-	484
	Rep	1796	350	519	728	-	67
	-	193384	27058	40071	105182	-	12852
	All	202253	28176	42038	110040	-	13403
KBJ	Dem	6845	1240	626	1229	-	750
	Rep	1555	473	141	831	-	110
	-	261252	46109	25522	155699	-	33922
	All	269562	47822	26289	160759	-	34782

Table A2: Tweets collected for each nominee using automated TAGS software. Column labeled 'Ford Hearing' contains tweets collected during the additional day added to Kavanaugh's hearing for testimony from Dr. Blasey-Ford.

AFINN Sentiment Dictionary

While the Finn (2011) Lexicon is well suited to our purpose, it is not perfect. Among the 2400 words in the Lexicon, many of them would indicate sentiment in a normal conversation, but during the confirmation process of a Supreme Court justice, might mean something very different. Take the word "justice" itself. In most situations, justice is a positive word indicating something fair, correct, or right has occurred. Indeed, AFINN gives justice a positive score of 2. However, justice in our situation would much more often refer to a name or a position (e.g. Justice Ginsburg), which would not indicate sentiment at all.

As such, we began by excluding the following words: justice, alive, dead, death, landmark, lawsuit(s), legal, questioned, questioning, reach*, suing, united, verdict(s). We also excluded any two-word phrases because we tokenize our sentiment analysis into unigrams (single word phrases) so two word phrases in the dictionary would put the whole sentiment score on just the first word. Finally, we excluded several words which were highly likely to be used sarcastically and whose valence we could not be sure of including: lol, ha, haha, hahahaha. These words are all categorized by the dictionary as very positive but often used very negatively on Twitter. Rather than making assumptions about the valence of these words, we simply did not count them.

A Replication: Does Twitter Data Align with Nominees' Public Opinion Polling?

Here we examine whether tweets about nominations and confirmations bear any relation to public opinion polling. We believe this is an important addition in thinking about not only our results, but in the utility of analyzing partisan behavior on Twitter. Several studies have utilized Twitter data to monitor public opinion, with one study even showing that Twitter sentiment mirrored the ebbs and flows of President Obama's Gallup job approval rating (O'Connor et al. 2010). To explore how well our Twitter data stack up to this seminal research, we begin by collecting a baseline of public attitudes toward each of the nominees represented in our Twitter data. We do so by using a series of Gallup polls that ask: "As you may know, [nominee] is a [current position] who has been nominated to serve on the Supreme Court. Would you like to see the Senate vote in favor of [nominee] serving on the Supreme Court, or not?" This question has been asked since Robert Bork's nomination, uses consistent wording, and there are eight such polls for our three nominees (five for Kavanaugh, two for Barrett, and one for Jackson). Aggregate support is measured as the percent of respondents who selected "yes, vote in favor" for each nominee.

With this baseline established, we next follow O'Connor et al. (2010)'s coding to create a positive-to-negative tweet ratio for each event. In this configuration, our sentiment ratio variable grows as positive tweets outpace negative tweets, making it comparable to Gallup's percent of respondents who support the nominee. In order to statistically analyze sentiment vis-a-vis public polling data, we orient both data sets by date and then carry the most recent sentiment ratio forward to align with the nearest future Gallup poll. For example, Brett Kavanaugh's initial confirmation hearings took place from September 4th-7th of 2018 and we gathered tweets for this time-frame, however the nearest future Gallup poll was administered September 10th-16th. The sentiment ratio calculated for the hearings was therefore carried forward to align with the Sept. 10-16 polling period. Utilizing this method, sentiment and public approval are correlated at $r = 71.1\%$, nearly the same level reported by O'Connor et al. (2010) ($r = 73.1\%$). Further, if we perform an ordinary least squares regression, using only sentiment to predict the nearest future Gallup poll, sentiment is positively signed and statistically significant, accounting for over 50% of the variation in public support ($t = 2.47; p = 0.048; R^2 = 0.505$). These results become even more sharp if we restrict our analysis to only consider partisans.[1] In this configuration, our correlation statistic rises to $r = 81.9$ while sentiment accounts for over 67% of the variance in OLS regression ($t = 3.50; p = 0.013; R^2 = 0.6709$). To provide a better idea of how these two measures relate to one another, Figure A1 depicts both sentiment and public support for the nominees; our variables are rescaled between 0 and 1 for ease of comparison.

At first glance, the two measures behave quite consistently with one notable difference. While sentiment is lower than and tends to track with public support for Barrett and Jackson, we see this relationship inverted for a portion of Kavanaugh's proceedings. Although both sentiment (dashed line) and Gallup's public support (solid line) trend downward between Kavanaugh's announcement and hearings, support subsequently peaks while sentiment moves to its lowest point. This low-point for sentiment is measured September 27th and 28th of 2018 to coincide with Dr. Christine Blasey Ford's testimony that Brett Kavanaugh sexually assaulted her at a high school party. Meanwhile, the Gallup data were collected September

1. Specifically, we determine the number of tweets per nominee and event generated by Republicans and Democrats and then sample from the larger group a number equal to the smaller. For example, if we captured 600 tweets by Democrats and 400 tweets by Republicans during Amy Coney Barrett's Hearings, we randomly sample 400 Democratic tweets, thus ensuring neither partisan group is over-represented when calculating mean sentiment. We additionally compute and average 100 iterations of each random draw to ensure no single sample will skew results.

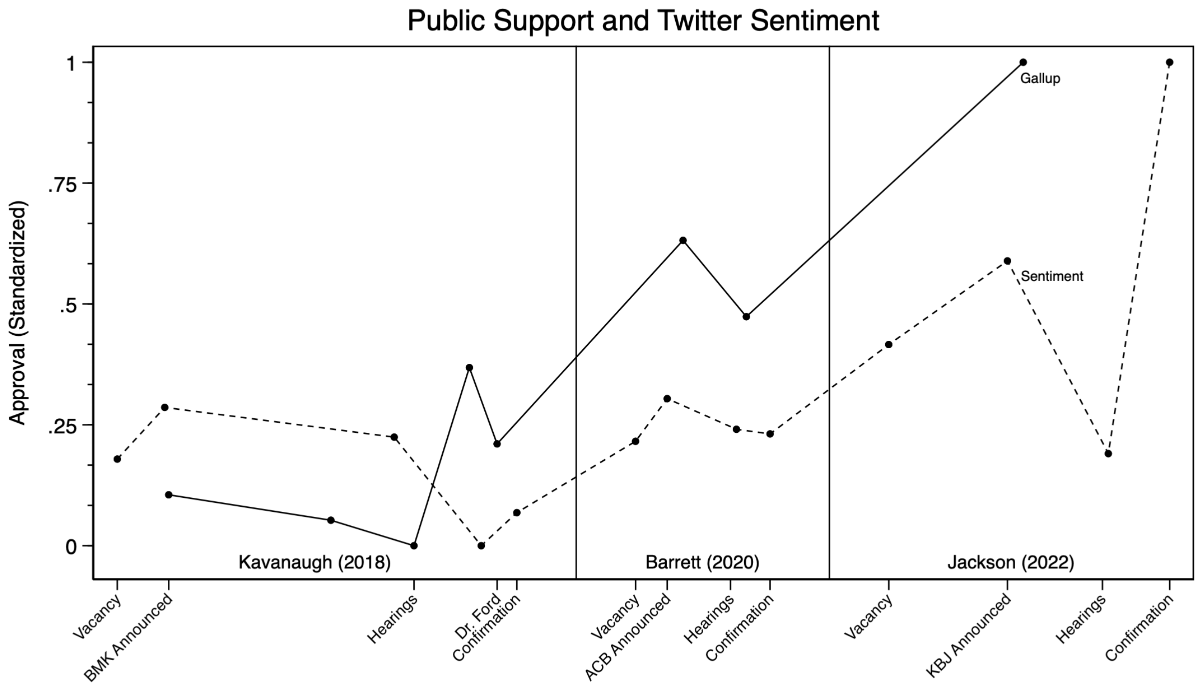

Figure A1: Public support (solid line) and sentiment ratio (dashed line) for Kavanaugh (first section), Barrett (middle section), and Jackson (right section).

24th - 30th, meaning at least some portion of those data were collected *before* Dr. Blasey Ford's testimony. We see a similar trend repeated during Ketanji Brown Jackson's confirmation process; Gallup finished its final poll on March 18th, 2022 – three days *before* the start of Jackson's Senate hearing and therefore did not capture attitudes about the hearing itself. Despite this divergence and similar to prior studies utilizing microblogs, our data seem to track – generally – with public opinion polling, both statistically and visually. As noted above, we don't go so far as to claim Twitter data *are* public opinion, but this analysis and prior studies demonstrate they're also not entirely disconnected from mass polling results. We do, however, believe that this lends validity to thinking of Twitter data as providing critical insights to the discussions and sentiments expressed during political events.

Difference of Means Tests

Here we present a series of t-tests to compare types of sentiment (overall, positive, and negative) across our partisan groups. We compare the sentiment of partisans who align with the nominee to those who do not align with the nominee ideologically by first excluding the vacancy stage from our $t-tests$ as this allows us to isolate reactions to the nomination process rather than reactions to the vacancy. Across all three types of sentiment (overall, positive, and negative), we find statistically significant differences where partisan 'winners' tweet more positively than partisan 'losers' ($t = -6.13, p < .01$; $t = -8.07, p < .01$; $t = -2.42, p < .01$, respectively). We proceed to carry out similar t-tests at the vacancy stage. The same idea holds; partisans should tweet more positive things about a vacancy on the Court if the presiding president is of their same party. Similar findings emerge. At the vacancy stage across all three types of sentiment (overall, positive, and negative), we again find statistically significant differences where those who stand to benefit from a vacancy tweet more positively at that stage ($t = -4.60, p < .01$; $t = -2.97, p < .01$; $t = -4.06, p < .01$, respectively). Collectively, these sentiment data tell us that partisans can and do view confirmation processes through their partisan lenses.

Topic Modeling

Gathering Additional Tweets for Partisan Users

Topic modeling generally works better with more text – there is, of course, a ceiling to this principle, but when working with short-texts such as tweets, we're not meeting that ceiling (Sbalchiero and Eder 2020). Amazon Web Services' Developer Guide, for example, recommends using at least 1,000 documents when utilizing LDA models.[2] Since we define each tweet as its own document, and given that our initial crop of tweets did not include 1,000 partisan-identified tweets per event, we made the decision to gather more tweets from

2. https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html

our hand-identified partisan users. To do so, we utilized Twitter's API v2 with an academic account to create queries using our original search terms, our original search time-periods, but exclusively searching user accounts that were coded for partisanship. We limited this secondary search to collecting only 10 tweets per user per event to ensure that no single user was able to dominate a topic. Although we did not use these tweets in our sentiment analysis (since they weren't part of the initial random sample we use to compare to public opinion polling), we note that including them does not change the results of those analyses.

		Tot.	Vacancy	Announc-ement	Conf. Hearing	Ford Hearing	Final Vote
BMK	Dem	15253	2475	2079	4121	4117	2461
	Rep	5040	893	877	1188	1057	1025
ACB	Dem	13615	3606	2668	6127	-	971
	Rep	4570	1415	1151	1617	-	255
KBJ	Dem	10250	1700	851	6300	-	1399
	Rep	2650	680	220	1476	-	274

Table A3: All partisan tweets collected for each nominee using automated TAGS software and supplemented by API v2 queries by partisan user. Column labeled 'Ford Hearing' contains tweets collected during the additional day added to Kavanaugh's hearing for testimony from Dr. Blasey-Ford.

LDA Models & Comparison Method

When initially exploring these data with topic modeling, we did consider and test the STM model on a sample of data (4000 random tweets from the Kavanaugh hearings). Using the package's settings to auto-detect the ideal K yielded 59 topics that over fit the sample. We recovered similar results when specifying the model with partisanship as a content variable, as a prevalence variable, and as both, and when replicating the LDA analysis we present in our main manuscript's Table 1. Subsequently examining the topics produced by the stm package after manually choosing k based upon diagnostic metrics solidified our belief that the STM model wasn't best-suited to our data. Anyone with even a cursory understanding of the proceedings for our three nominees would expect topics relating to sexual assault for the

BMK hearing, women's rights for the ACB hearing, and sentencing for child pornography in the KBJ hearing. Most surprisingly, the topics produced for Kavanaugh's hearing did not include any mention of sexual assault, Dr. Ford, or the popular hashtags #believesurvivors and #metoo. We could, of course, 'fix' this by choosing a different K such that this topic *is* represented, however selecting K in order to yield a specific topic introduces far too much user bias.

In short, the stm package seemed to want to over fit our data, producing auto-fit results and diagnostic metrics that skewed toward higher Ks – and that in the Kavanaugh dataset yielded especially perplexing results. We wonder if this might be because, when compared with LDAs, the STM tends to produce topics with higher metrics for topic exclusivity, but lower semantic coherence scores (Roberts et al. 2014, Appendix, 19). This bias toward exclusivity seems less ideal for our data which are collected around specific keywords and modeled over a single confirmation process. This means our data are already somewhat structured (this helps mitigate sparsity issues that can arise when using LDAs with short texts), and probably *should* have overlap between the topics. While the stm package is a powerful tool that allows direct integration of covariates (that we're eager to use in other contexts!), the vanilla LDA seems more suited toward our particular data.

Prior to implementing an LDA model, we first tokenize and lemmatize our tweets. These processes remove any emojis, foreign characters, or stopwords, and then break down words to the root level. For example, a tweet that initially reads, "judge kavanaugh did very well today but satan never gives up and this situation has been laden with spiritual warfare. keep praying!!" becomes "judge kavanaugh today satan situation laden spiritual warfare praying" after tokenizing and lemmatizing. This process ensures that topics aren't overrun with common words such as 'the' and 'but' and that word roots are prioritized over cognates.[3]

After this pre-processing step, we download Lu, Henchion, and Namee (2019)'s suite of

3. For a discussion of tokenizing, lemmatizing, and stopwords, see (Albalawi, Yeap, and Benyoucef 2020, 9).

Python scripts from the Github repository referenced in their article.[4] In simple terms, Lu, Henchion, and Namee (2019)'s comparison method follows three steps: 1) combine sets of documents from two or more unique sources and run an LDA model on those combined documents using gensim's ldamodel module; 2) select a model based upon the researcher's knowledge of the documents and the model coherence scores; 3) recover the top n discrete topics by source.

To accomplish step one, we download Python's ldamodel module from gensim,[5] and run an initial model on our combined corpus using Lu, Henchion, and Namee (2019)'s pre-set parameters. Since we aren't using our LDA models to classify an untrained set of data, we use all tweets by users that had been hand-coded for partisanship, as seen in Table A3. To accomplish step two, we next run several iterations of models searching for meaningful topics with an optimized coherence score (Albalawi, Yeap, and Benyoucef 2020). To do so, we varied the following parameters:

- topics: We preliminarily explored number of topics, K, in increments of 10 from 10-100. This suggested a more targeted approach by 4s from 4-32, and then then fine-grained iterations from 8-20 by 1s. We narrowed in on the upper-range of our K-search by letting the diagnostic metrics objectively guide guide us, however we did choose a lower-limit (8) based upon knowledge of our data. This method was adopted to ensure there were both sufficient topics such that some specificity and difference could be detected between partisan groups, but not so many topics that they represented only a handful of tweets apiece (Sbalchiero and Eder 2020).
- passes: by 5s, from 10-60.
- iterations: by 10s, from 40-100.
- random state: 42, 1984, 2000.

4. https://github.com/GeorgeLuImmortal/topic-based_corpus_comparison/blob/master
5. https://radimrehurek.com/gensim/models/ldamodel.html

After identifying that somewhere between 10-20 topics was ideal for meaningfulness, we ultimately selected our final models based upon computed coherence score. This was done to ensure researcher bias (i.e. what we wanted to see in the results) didn't impact our model choice. Graphic depictions of the models for each nominee are found in the attached files named topics_NOMINEE. The left-hand side of each graphic presents a spacial mapping of each topic, where size of the topic's circle represents its prevalence in the corpus, and distance between circles indicates the relationship (or lack thereof) between topics. The right-hand side of each graphic shows the top 30 words relevant to each topic. Note that this graphic is interactive, so clicking on a circle in the left panel will dictate which words appear on the right. Similarly, clicking on a word on the right side will highlight that word's prevalence in the left panel's topics.

Finally, we complete step three of Lu, Henchion, and Namee (2019)'s comparison process by running their topic_based_cc.py script which produces a list of the top n topics, ranked by most uniquely belonging to a particular source. To produce this list, Lu, Henchion, and Namee (2019)'s method fits a single LDA model and then applies one of four user-selected divergence metrics "directly to topic membership vectors rather than applying [them] to word distributions between topics" (p. 3). We choose the Chi-squared metric due to its comparative performance (p. 8), and pair that with the weighted sum aggregation option since we do not have even numbers of tweets from identified Democrats and Republicans (see Table A3).

While more technical aspects of the method are available in Lu, Henchion, and Namee (2019)'s paper, their code determines how much each source (for our data, Republicans or Democrats) contributes to each topic. They do this by taking the topic proportions assigned to each document, and averaging these by each source. Topic 1, for example, might have an average proportion of .02 for Democrats and .035 for Republicans with varying distributions. These means and distributions of topic proportions are then compared using a discrimination metric in order to calculate divergence.

Since we identified our sources as Democrats and Republicans, the output allows us to ID the topics that are most unique, and then attribute those topics to rhetoric originating from partisan users. To be clear, the comparison process does not change the topics found by the original LDA model – it simply allows us to better understand the predominant source of those topics.

		Vacancy	Announcement	Conf. Hearing	Ford Hearing	Final Vote
BMK	**Dem**	trump republican roevwade nominee vote save flake right voter overturned	president kavanaugh judge investigation trump law justice judgekavanaugh criminal sitting	kavanaugh senate lied document email booker stopkavanaugh republican time perjury	kavanaughhearings believesurvivors kavanaugh metoo graham woman lindsey lindseygraham white men	kavanaugh ford collins woman senatorcollins republican dr investigation susan susancollins
	Rep	justice abortion liberal court retirement majority kennedy right conservative kagan	news live fox pick foxnews trump protest nominee protester tonight	hearing kavanaugh confirmation judge confirmationhearing judgekavanaugh democrat live protest abortion	kavanaugh judge judgekavanaugh justice democrat trump high school kavanaughhearings hearing	kavanaugh ford christine graham story lindsey blasey christineblasey justice washington
ACB	**Dem**	woman right justiceginsburg rest rip life fight year peace justice	republican trump care right biden law know vote election aca	vote trump care republican election aca senate american health healthcare		trump republican state right want united unitedstate american court obamacare
	Rep	justiceginsburg trump replace vote cruz justice ted replacejusticeginsburg dead tedcruz	trump judge barrett nominee judgebarrett president court term vote conservative	control judicial scalia birth senator hirono hate birthcontrol explain love		vote senate republican committee judiciary nomination judge confirmation barrett trump
KBJ	**Dem**	mcconnell republican mitch seat biden pick nominee year block mitchmcconnell	judge jackson biden nominee president judgejackson justice republican court brown	hearing confirmation confirmationhearing trump graham kavanaugh lindsey justice lindseygraham thomas		black woman justice blackwoman senate judge united unitedstate state jackson
	Rep	biden pick race justice joe based joebiden breyer judge president	biden pick nominee republican graham bidenpick trump lindsey decision ryan	biden great case woman define husband recuse american watching pick		black republican vote scott justice tim man think child democrat

Table A4: Most divergent topic by partisan user and by nominee for separate stages of the nomination and confirmation process. Gray cells represent de-facto policy losers.

References

Albalawi, Rania, Tet Hin Yeap, and Morad Benyoucef. 2020. "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis." *Frontiers in Artificial Intelligence* 3. ISSN: 2624-8212.

Finn, Årup Nielsen. 2011. "A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs." In *ESWC 2011 Workshop on 'Making Sense of Microposts',* 93–98. CEUR Workshop Proceedings, May.

Lu, Jinghui, Maeve M. Henchion, and Brian Mac Namee. 2019. "A Topic-Based Approach to Multiple Corpus Comparison." In *Irish Conference on Artificial Intelligence and Cognitive Science.*

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. "From tweets to polls: Linking Text Sentiment to Public Opinion Time Series." In *International AAAI Conference on Web and Social Media,* 122–9.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58 (4): 1064–82.

Sbalchiero, Stefano, and Maciej Eder. 2020. "Topic modeling, long texts and the best number of topics. Some Problems and solutions." *Quality & Quantity* 54 (4): 1095–1108. https://doi.org/10.1007/s11135-020-00976-w. https://doi.org/10.1007/s11135-020-00976-w.