SMaPP Global Special Issue - Original Research

SAGE Open
January-March 2019: 1–13
© The Author(s) 2019
DOI: 10.1177/2158244019828850
journals.sagepub.com/home/sgo
SAGE

Why Keep Arguing? Predicting Engagement in Political Conversations Online

Sarah Shugars[1] and Nicholas Beauchamp[1,2]

Abstract

Individuals acquire increasingly more of their political information from social media, and ever more of that online time is spent in interpersonal, peer-to-peer communication and conversation. Yet, many of these conversations can be either acrimoniously unpleasant or pleasantly uninformative. Why do we seek out and engage in these interactions? Who do people choose to argue with, and what brings them back to repeated exchanges? In short, why do people bother arguing online? We develop a model of argument engagement using a new dataset of Twitter conversations about President Trump. The model incorporates numerous user, tweet, and thread-level features to predict user participation in conversations with over 98% accuracy. We find that users are likely to argue over wide ideological divides, and are increasingly likely to engage with those who are different from themselves. In addition, we find that the emotional content of a tweet has important implications for user engagement, with negative and unpleasant tweets more likely to spark sustained participation. Although often negative, these extended discussions can bridge political differences and reduce information bubbles. This suggests a public appetite for engaging in prolonged political discussions that are more than just partisan potshots or trolling, and our results suggest a variety of strategies for extending and enriching these interactions.

Keywords

politics, social media, interpersonal communication, deliberation, polarization, natural language processing, topic models, sentiment

Introduction

Digital communication plays a critical role in our political infrastructure. Online platforms have expanded the reach of "kitchen table conversations," as people increasingly turn to social media as a primary news source (Bakshy, Messing, & Adamic, 2015; Lee & Ma, 2012; O'Connor, Balasubramanyan, Routledge, & Smith, 2010) and elected officials use digital channels to communicate with their constituents (Farina, Epstein, Heidt, & Newhart, 2013; Kavanaugh et al., 2012). Such interactions are often modeled as one-shot games or as evidence of long-term links in a social network (Feng & Wang, 2013; Myers & Leskovec, 2014). Yet, much online activity consists not of single-shot, unidirectional interactions with elites, but repeated interactions among peers. These iterated interactions—conversations—have important implications for political theory: while conventional wisdom claims that brief social media interactions have little effect on subsequent online behavior, a number of recent experiments have shown modest but real effects of single-shot interactions (Friggeri, Adamic, Eckles, & Cheng, 2014; Munger, 2017). Deliberative theory suggests that repeated interpersonal interactions where individuals engage in extended conversation may have even more substantial effects (Axelrod, 1987; Bednar & Page, 2007; Habermas, 1984).

In this article, we focus not on persuasive outcomes but on the more fundamental question of what leads people to engage in extended online conversation and argument in the first place. Existing work in this area has generally had a more practical bent, focusing on tweet- or conversation-level recommendation and aiming to predict user interest in conversation threads to better curate and recommend targeted content. Such work has looked at user engagement in various forms of online conversation (Chen et al., 2012; He & Tan, 2015; Vosecky, Leung, & Ng, 2014; Yan, Lapata, & Li, 2012), as well as via retweeting (Feng & Wang, 2013; Hong, Doumith, & Davison, 2013) and re-entry back into existing

[1] Network Science Institute, Northeastern University, Boston, MA, USA
[2] Department of Political Science, Northeastern University, Boston, MA, USA

Corresponding Author:
Sarah Shugars, Network Science Institute, Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA.
Email: shugars.s@northeastern.edu

conversations (Backstrom, Kleinberg, Lee, & Danescu-Niculescu-Mizil, 2013). Our work here is closest to the latter: We are more interested in the extended dynamics of conversation, particularly the decision to re-engage or exit, than in the initial decision to interact. We focus on users who have already made that initial participation, and seek to understand and predict whether and when they re-engage based on user, tweet, and thread-level features.

While ongoing deliberative conversation is the substantive focus here, this framing also turns an impossible problem—predicting initial responses to a tweet out of the entire pool of twitter users—into a practicable prediction task—predicting re-participation of users whom we know are already part of a conversation. This approach also conditions out the even harder problem of explaining the origins of an initial tweet or conversation, particularly given the immense variety of motivations behind those first moves. Instead, we focus on existing conversations—at least a first move followed by a response—and model the processes that lead to extended and branching conversations among existing participants. Twitter might seem less suited to such models than traditional online forums, but Twitter in fact produces immense quantities of impromptu extended, branching conversations, and by focusing only on re-entry by existing participants, we can study what causes individuals to continue an argument or drop out, bracketing the question of initial engagement.

By conditioning on existing user interaction, we aim to get more deeply at the question of why people bother arguing online. What brings them back to a repeated argument? What factors contribute to an individual returning to or abandoning an argument? While in face-to-face settings, social etiquette suggests that a comment will most likely be greeted with a response, there is no a priori reason to expect a response to the vast majority of online posts. While we expect to find many types of conversations occurring online, we might expect that more extreme content (positive or negative) will increase engagement, as trolls successfully incite arguments and partisan allies reinforce each other's positions (Cheng, Danescu-Niculescu-Mizil, Leskovec, & Bernstein, 2017). Between the extremes lies a more productive and interesting mode of engagement: true deliberative argument, in which participants exchange content in a genuine attempt to persuade or inform. Such behavior is not as uncommon as skeptics might assume, and is prevalent in knowledge-sharing platforms such as StackOverflow, Yahoo! Answers, and other such forums, where users may be motivated to some degree out of a general sense of community (Adamic, Zhang, Bakshy, & Ackerman, 2008; Anderson, Huttenlocher, Kleinberg, & Leskovec, 2012; Oktay, Taylor, & Jensen, 2010) even as they argue over better or worse solutions to shared problems.

We find evidence for all of these behaviors in our data and, in particular, show that while many of these engagements are negative, conversations often cover a range of emotions and go on far longer than a single-shot attack or mutual trolling might suggest. While we leave for later the ultimate question of persuasive effect, we establish here that even a medium as apparently unpromising as Twitter is full of complex, extended political conversations, and that individuals' decision to repeatedly re-engage in those conversations is surprisingly systematic.

Related Work

Much of the theoretical work around questions of conversational dynamics has been done within the literatures of deliberation and persuasion. While both these approaches focus their attention on back-and-forth conversations, they vary in their characterization of those conversations. The persuasion literature looks broadly at how people convince others and "win" arguments, while the deliberative ideal imagines thoughtful participants reasoning together to generate public opinion centered on the common good (Cohen, 1989; Habermas, 1984). "Reasons" may constitute factual arguments or emotional appeals (Mansbridge, 2015), but ideal deliberation is often taken to be free of persuasion, coercion, or other forms of instrumental action (Habermas, 1984). Contra persuasion models, ideal deliberators should engage in rational speech acts—aiming to honestly express themselves and truly trying to understand the other. Huckfeldt, Mendez, and Osborn (2004) argue that ideal citizens "are those individuals who are able to occupy the roles of tolerant gladiators—combatants with the capacity to recognize and respect the rights and responsibilities of their political adversaries" (p. 91) If political debate serves to sharpen our own understanding and build our collective knowledge, then we owe it to our interlocutors to press them on their positions, to find the holes in their armor and encourage refinement of beliefs. The process of debate makes us all better—thus allowing tolerant gladiators to walk away as friends. Citizens who silence their discussants, seek to coerce others, are easily persuaded by false beliefs, or who otherwise refuse to engage in rational argument, therefore, do a disservice to themselves and to their communities.

Experience tells us, however, that such a lofty deliberative ideal is rarely met in political conversation. Sunstein (2002) advances the "law of group polarization," finding through numerous empirical studies that "deliberation tends to move groups, and the individuals who compose them, toward a more extreme point in the direction indicated by their own pre-deliberation judgments" (p. 175). Sunstein argues this polarization is the natural result of the social context, which serves as a significant driver of individual actions and opinions. Hearing friends express a view makes a person socially inclined to express the same view. In other words, deliberating groups tend toward extremism in the direction of the pre-deliberation median because nobody wants to take the social risk of expressing an unpopular view. Sanders (1997) similarly argues that the broader context of power dynamics

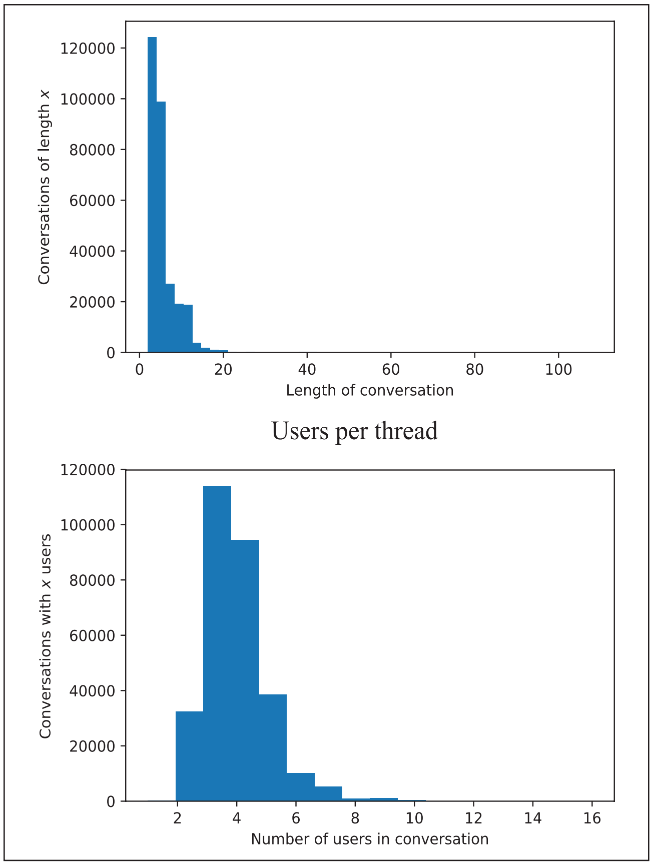

Figure 1. Distributions of length and users by thread.

frequently has a debilitating but under-recognized effect on deliberation, as marginalized individuals feel silenced and unable to share their true opinions. Importantly, the majority of participants may mistakenly assume that such power effects are negligible if "deliberation appears to be proceeding."

Another line of work has tackled conversational dynamics from the perspective of the data-processing problem of platform curation, for example, trying to predict which posts will be popular for the purpose of highlighting those posts for users. Much of this work focuses on post-level engagement, predicting engagement as a function of topics (Hong et al., 2013) or social network structures (He & Tan, 2015; Pan, Cong, Chen, & Yu, 2013). Much of this work has considered "popularity" as a raw aggregate of engagement with an initial post, finding, perhaps unsurprisingly, that the popularity of a user's past content is a strong predictor for the popularity of their future content (Artzi, Pantel, & Gamon, 2012). Backstrom et al. (2013) break the task into related subtasks: length prediction and re-entry prediction. Intuitively, these subtasks indicate distinctive types of threads: threads which are long because a high number of users chime in a small number of times—to offer congratulations or condolence, for example—while other threads are long because a small number of users contribute a large number of times in a back-and-forth conversation. Supporting this theory, Backstrom et al. (2013) find that the number of distinct users in long threads follows a bimodal distribution. Using data from Facebook and Wikipedia, Backstrom et al. (2013) find the identities of recent commenters is most predictive of conversation re-entry.

This lattermost line of work is largely atheoretical and not particularly concerned with normative issues. While the current study borrows many of their methods, we are also fundamentally interested in the dynamics of online conversation from a deliberative perspective. Thus, we are interested less in conversation recommendation or modeling engagement in conversations per se, and more focused on how individual speech acts (tweets) lead existing discussants to re-engage with each other or abandon a conversation. Regardless of the outcome of a conversation, it is important to understand what sustains conversations—particularly acrimonious ones—and keeps mutual opponents or supporters engaged with each other. As we will see, this engagement can take more or less productive forms, but simply understanding the deliberative dynamics is an important first step.

Data

For this study, we collected a corpus of 7,053 Twitter conversations during the month of October 2017 seeded by tweets with the keyword "Trump." For each tweet discovered through keyword search, we extract the entire conversation tree of preceding and following replies, if there is one. Such trees may be composed of multiple branching threads, each connecting to the same root tweet.

We then used the Twitter API (Application Programming Interface) to retrieve tweet metadata for each tweet in the conversation tree. We discard trees which have no conversations longer than a minimum of three exchanges or in which tweets have been deleted, as metadata for those tweets cannot be retrieved.

Our full corpus contains 7,053 conversations comprised of 63,671 unique tweets. The distribution of thread length is heavy-tailed: by construction, the minimum thread length is 2, while the longest thread contains 108 tweets. The average length of a thread is 5.6 tweets with a standard deviation of 4.1, and the mean number of unique users in a conversation is 3.7, with a standard deviation of 1.2. The distributions of length and users by thread can be seen in Figure 1. Furthermore, as we might expect from social media engagement, responses tend to occur within a relatively compressed time period. Just under half (40%) of the tweets in our sample are posted within 5 min of the tweet which proceeds it in the conversation. About three quarters (70%) are posted within 1 hr, and nearly all (95%) take place within a day. The cumulative distribution of inter-event time—that is, the number of hours between tweets—can be seen in Figure 2.

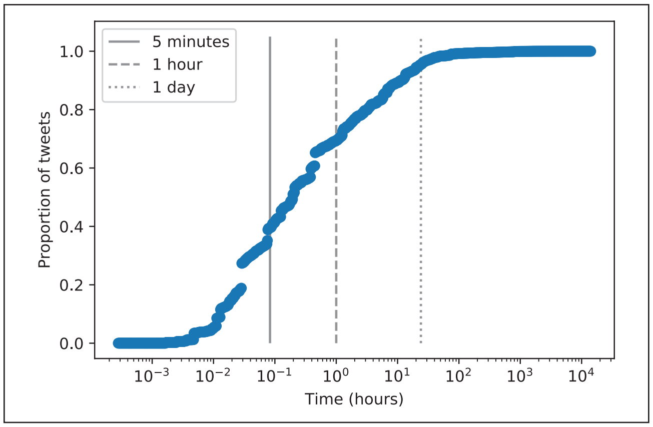

Figure 2. Cumulative distribution of the time taken to reply to a tweet for those tweets with replies.

Model

Our fundamental question is why someone continues to engage in (or stops engaging in) an online conversation. What makes a person participate in or abandon a discussion? We examine this question by modeling conversation as an interlaced exchange between two or more participants. For a tweet observed at time step t, we wish to predict whether existing members of the conversation, as defined below, will respond or not respond to that tweet at time step $t+1$.

While anyone in the Twitter universe may conceivably reply to a tweet, our interest is in modeling the actions of those who are already part of a conversation in a loose sense: first, because this makes the prediction problem practical and, second, because it allows us to model engagement in dialogue, not just taking potshots on a microblog. Furthermore, as many threads on Twitter are initiated by entities unlikely to participate in conversation—such as corporations, celebrities, politicians, and bots—we take our pool of candidates to be users who have already *responded* at least once in a given thread, ruling out the user who initiated the thread unless they also replied to another tweet in the conversation. We consider self-replies to be a continuation of a thought, and thus not a "reply" in the traditional sense. We therefore do not include the author of a tweet in the list of candidates who may respond to that tweet.

Our predictive model is structured as follows: For conversation j, at every time step $t>2$, we construct a candidate list of active participants who might respond to the current tweet. Those candidates who do reply at time $t+1$ are assigned an observed outcome of 1 while all potential respondents who do not choose to reply are assigned an outcome of 0. Note that multiple users may respond directly to a single tweet. In this analysis, we focus on the temporal sequence of replies and use t as an index of that temporal order. Users return to Twitter on their own schedules, generally due to exogenous constraints on their free time, and have the opportunity to respond to any available tweet when they do. Only two thirds (66%) of users in our dataset make all their comments within an hour of their initial activity, suggesting that it is common for users to engage in conversation over multiple Twitter sessions. While we may generally expect returning users to respond to the most recent tweet, conversational engagement need not follow this temporal order, and users in our dataset seem to flit between the threads of a conversation tree, going back to respond to earlier tweets nearly half (43%) of the time. For all these reasons, measuring t via clock time intervals, for example, within a Poisson framework (George & Kibria, 2011; Shen, Wang, Song, & Barabási, 2014), would overlook this exogenously constrained, bursty, and often non-sequential reply behavior, whereas using temporal sequence order allows us to relax this assumption and look at the conversational points at which a user re-engages with the understanding that nonengagement could be due to any number of factors. We do, however, continue to use tweets' time stamp for the calculation of certain features, specifically the number of likes, retweets, and comments visible at the time of a candidate's reply, as well as to control for time-of-day patterns that affect when users generally engage with Twitter.

Table 1. Example Conversation Flow Between Participants A, B, and C.

t	Conversation order at time t	Candidates	Observed at $t+1$
3	$A \rightarrow B \rightarrow A$	B	$B=0$
4	$A \rightarrow B \rightarrow A \rightarrow C$	A, B	$A=0; B=1$
5	$A \rightarrow B \rightarrow A \rightarrow C \rightarrow B$	A, C	

Table 1 illustrates a conversation thread from time step $t=3$ to time step $t=5$. At each step, we show the potential candidates for re-entry, and the outcomes associated with that time step; the number of observations at each time step is equal to the number of potential respondents; only those candidates who respond are scored as a 1, with the rest assigned a 0 for that time step in that conversation.

For our dataset of 7,053 conversations, this results in 1,016,492 total observations, with 110,035 observed instances of 1 (candidate users who responded) and 906,457 observed instances of 0 (candidate users who did not respond). This gives us a baseline prediction accuracy of 89% if we guess that all candidates never respond.

Features

We expect a user's tendency to reply to a conversation to be influenced by a number of factors and their interactions. Previous work (Artzi et al., 2012; Backstrom et al., 2013; Feng & Wang, 2013; Hong et al., 2013) has generally focused on predicting conversation-level engagement (i.e., whether a user participates anywhere in a conversation) and has, therefore, primarily focused on the candidate user who might reply as well as features of the overall conversational thread. As we are interested here in the more specific problem of predicting the points at which an existing participant replies or does not reply, we further include features related to the author who might receive a reply, as well as the tweet that may be replied to.

This gives us three sets of features and related hypotheses which we discuss in the follow subsections:

> **Hypothesis 1 (H1a-H1e):** Candidate and recent tweet features: This includes features related to the candidate user's activity (H1a-H1c) as well as features of that candidate's previous tweet in the thread (H1d-H1e).
> **Hypothesis 2 (H2a-H2b):** Conversation thread features: Features related to thread length and engagement (H2a-H2b).
> **Hypothesis 3 (H3a-H3e):** Author and current tweet features: Includes features of the author who may receive a reply (H3a-H3b) as well as the tweet at time t which may be replied to (H3c-H3e).

Candidate and Recent Tweet Features

At the most basic level, we would expect that active users will be more likely to reply at any given point in the conversation (H1a). The most readily available measures of engagement for a user are the number of others they follow (`following count`), the number of followers they have (`follower count`), the total number of tweets they have posted (`statuses count`), the number of favorites or likes they have given (`favourites count`), and whether they have a `verified` account.[1] While measures of activity such as the number of users they follow, number of tweets, and number of favorites given would all presumably have a positive effect on reply probability, measures of popularity such as the number of followers or being verified could possibly have the opposite effect, as more popular users may be less likely to enter a scrum with the hoi polloi (H1b).

At the level of thread–user interactions, we would also expect that a user is more likely to reply as a function of how engaged they have already been in the conversation (number of replies) up until that time, and presumably less likely to respond the longer it has been since their last comment (H1c). We include two features to capture this dynamic: a binary variable `prev response` which indicates whether the current tweet t was in response to this candidate user, and `time since prev` which provides a raw count of how many time steps it has been since the candidate's last comment. Similarly, we would expect that a candidate respondent will be more likely to re-engage if their most recent tweet in the conversation received positive feedback, as measured by the number of favorites, retweets, and replies the candidate's previous tweet received (`favorite count`,[2] `retweet count`, and `reply count`, respectively; H1d).

We also examine content-level characteristics for candidates by evaluating the topical distribution and emotional valence of their most recent tweet in the conversation prior to time t. We describe these features in detail in "Current Tweet and Author Features" section. We expect that candidates whose previous tweet was negative or more extreme in its valence are more likely to keep a conversation going (H1e). In addition, comparing the topical content of a candidate's previous tweet with the content of the current tweet, for which we are predicting response, allows for a measure of interest similarity between the two users. After constructing topic vectors, as described in "Current Tweet and Author Features" section, for a candidate's previous tweet and another user's current tweet, we calculate the absolute and squared euclidean distance between the two vectors. We treat this as an inferred measure of ideological `difference`, indicating whether conversations are likely to stay within topic, or alternate between users.

Conversation Thread Features

Following recent work in this area (Artzi et al., 2012; Backstrom et al., 2013; Feng & Wang, 2013; Hong et al., 2013), we would expect various features of the conversation up until time t to affect user engagement. The length of the conversation thread (`thread length`), for instance, is a good indicator of the amount of interest the conversation has generated and, therefore, is expected to increase the probability of reply (H2a). This effect, however, may decrease with increasing thread length (H2b) as the conversation becomes more disjointed, unwieldy, or difficult to display via the Twitter interface. We therefore also include a quadratic term for thread length.

Current Tweet and Author Features

Our approach differs most significantly from past models in that we are interested not only in user engagement overall in a conversation but also in predicting the conversational points at which a user chooses to engage. To model this requires accounting specifically for the features of each tweet that may be replied to, as well as features of that tweet's author.

For author characteristics, many of the user features discussed in "Candidate and Recent Tweet Features" section may also influence the probability of an author receiving a reply. This may be mediated indirectly via the tweet, or directly in those cases where the author is known to the potential respondent. The author's activity levels, for instance, might be positively correlated with the likelihood of reply, indicating a tendency to produce more engaging tweets (H3a). Conversely, while popularity may decrease a candidate respondent's likelihood of replying, a tweet from a popular author may be more likely to receive a response (H3b).

Table 2. Top 10 Words for Each Topic.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
hope	love	people	Good	people	thought	puerto	true	news	live
hillary	sad	pr	Mayor	lol	evidence	rico	wrong	fake	usa
bot	big	money	Day	black	funny	years	obama	real	war
agree	yeah	power	God	white	russian	lies	people	time	matter
cnn	people	dying	Work	racist	means	people	president	flag	country
happen	dont	water	Supplies	point	food	understand	vote	protest	marathi
states	person	tax	Great	hate	read	party	thing	stand	tweeting
liar	blame	hurricane	Job	guy	act	white	shit	talking	class
argument	wow	taking	San	bad	facts	rich	care	watch	leader
clinton	pr	days	Juan	problem	helping	world	donald	anthem	place

In regard to the current tweet itself, there are many coarse structural (as opposed to content) aspects which may reflect latent characteristics of the tweet such as its general popularity or interest. Using tweets' time stamps, we calculate the count of the current tweet's favorites, retweets, and replies which were visible at the time of a candidate's response. We expect these features to generally have a positive effect on reply probability (H3c), although the first—favorites—may not have the opposite effect, if this action reflects silent agreement rather than a tendency to respond. A related measure of tweet "quality" is the ratio of retweets to replies, which we also include. We also include the length of the tweet and the device used to post it.

Because we would expect cyclical variation in activity, as users are naturally more likely to be active during certain hours of the day and on certain days of the week, we control for this tendency using cyclic transformations (Cox, 2006) of the day and hour at which a tweet was posted. These features are represented with the `xday, yday` and `xhour, yhour` features.

Finally, at the level of the current tweet's content, there are a wide variety of semantic and other linguistic characteristics that might increase the likelihood of reply (H3d). A tweet which mentions a large number of users may be more likely to elicit a response from those users or others; a larger number of hashtags may similarly increase the probability of response; and there is some evidence to suggest that tweets that include information such as a URL will be more popular as well (Bakshy, Hofman, Mason, & Watts, 2011).

At the level of sentiment and emotion, we hypothesize that users will be more likely to engage in conversations which are more emotionally extreme—whether participating in shouting matches of negative emotion, or vigorously reinforcing each other with positive emotion (H3e). We measure the emotional content of a tweet using several methods. These approaches use existing dictionaries to assign a valence score to each word and calculate the overall emotional value of a tweet as the average valiance of its component words. We capture a tweet's `sentiment` using AFINN (Nielsen, 2011), and its `valence` using the extended ANEW lexicon (Warriner, Kuperman, & Brysbaert, 2013). We also use VADER (Hutto & Gilbert, 2014) to more closely examine the negative or positive charge of a tweet. In all of these methods, lower scores indicate negative words while higher scores indicate positive words. VADER provides separate measures for valence along positive, negative, and neutral dimensions, as well as a compound score, which provides a single valence measure compiled from the three dimensions. Both AFINN and VADER were developed primarily for measuring sentiment in social media corpora. We also use ANEW (Warriner et al., 2013) to calculate arousal and dominance scores for each tweet. Arousal indicates the intensity of emotion, from calm to intense, while dominance scores indicate the degree of control, from vulnerable to powerful.

Finally, to capture higher level semantic content, we use Latent Dirichlet allocation (LDA) topic modeling (Blei, Ng, & Jordan, 2003) to identify topics in the corpus. In this model, each topic has an associated word distribution, and each document (tweet) has an associated topic distribution; by inspecting the former, one can discern the "meaning" of each topic, and by inspecting the latter, one can discern the topical focus of a tweet or, aggregated over all an author's tweets, the topical interests of that author. We pre-process tweets by removing punctuation, user handles, and standard English and Spanish stopwords—the latter because tweets in our corpus contain code-switching between English and Spanish. In addition, we treat "Trump" as a stopword for this corpus as it is the search term from which conversations were collected. Running LDA for 10 topics,[3] we take as features the topic distribution of the current tweet (i.e., 10 features) as well as the topic distribution of the most recent tweet in the conversation by the candidate user.

Table 2 shows the top 10 words associated with each of the 10 topics derived from our corpus of tweets. Note that topics are arbitrarily numbered and the labels presented here do not reflect a ranking. We can see that our collection of political tweets from October 2017 focused on stories such as emergency response in Puerto Rico (Topics 3, 4, and 7), National Football League (NFL) players kneeling during the national anthem (Topic 9), racism (Topic 5), and

Table 3. Response Predictors: Candidate Respondent.

	Coefficient	Significance after *p* correction		
		FDR	Cluster	FDR + Cluster
verified	–0.625	***	**	*
followers count	–54.294	***		
following count	–0.187	***		
statuses count	0.026	***		
favourites count	0.005			
comments count	–0.170	***		

Note. FDR = false discovery rate.
*p < .1. **p < .05. ***p < .01.

comparisons between President Trump and Democratic leaders (Topics 1 and 8).

Results

Prediction Accuracy

As responses are so rare, for a little under 90% of the observations, one would predict a response correctly by simply guessing a non-response for every possible respondent. This sets a relatively high baseline for prediction, but we find that using a straightforward logistic regression with the features described above significantly improves upon this baseline, achieving 94% out-of-sample accuracy[4] in predicting exactly who among the previous participants in a conversation will and will not respond to a given tweet.

Using a support vector machine (SVM; Boser, Guyon, & Vapnik, 1992), we are able to increase that accuracy to 98%, suggesting that there may be significant interactions and nonlinear effects among our features. Our SVM model is especially good at predicting nonresponses, erroneously predicting a response when the truth was a nonresponse only about 1% of the time. Conversely, as there are so few responses, we more often erroneously predict a nonresponse, getting about 13% of the true responses wrong, although that only amounts to another 1% of the total sample. In all, about half of our errors are false 0s and half are false 1s, showing that the model does a very good job overall of predicting both when people will choose to respond and when they will chose not to.

Tables 3 to 6 show the coefficients from the logistic regression model, as interpreting per-feature effects for SVMs is notoriously problematic. However, even for the logistic regression, identifying which features are "significant" is a non-trivial problem with so many features. With 1,016,49 observations, by most traditional measures of statistical significance, almost all of our features are statistically significant, regardless of the substantive magnitude of their effect. Even after multiple testing correction,[5] most coefficients are still significant.

However, in another sense, we do not have nearly as many observations as it may appear, as for any given tweet, very few choose to respond, and most responses are 0s. Furthermore, all tweet- and author-level conditions are shared across all the individuals who may or may not respond to that tweet. Thus, it makes sense to cluster errors at the current-tweet level, reflecting the fact that the number of observations with variation in tweet- and author-level features is far fewer than the simple count of observations would imply. After doing this, approximately half of our features lose their significance, and even more do so if we run false discovery rate (FDR) correction after error clustering. However, from a prediction point of view, this may be going too far, as our testing suggests that almost every feature—even if not significant by traditional statistical measures—does increase out-of-sample accuracy. This gap between the prediction and statistics literatures (Lo, Chernoff, Zheng, & Lo, 2015; Shmueli, 2010) goes beyond the scope of this article, so we present significance levels for all three corrections, and focus on the cluster-corrected version in most cases as being the most prevalent approach in the social sciences.

As in the previous section, we discuss the feature effects by category, as each category speaks to a different family of hypotheses. But it should be reiterated that Tables 3 to 6 all derive from the same single logistic regression, and are only broken up for convenience.

Response Predictors: Candidate Respondent

Table 3 shows results for features pertaining to the candidate respondent who has previously participated in the conversation and now may decide whether to respond to the current tweet or not. At the general level, we find that as expected in H1b, more popular users are less likely to re-engage even though they have done so already. Similarly, users are less likely to respond if they are verified or have more followers, although this effect is more fragile to cluster correction.

Interestingly, although it is also nonsignificant after clustered-error correction, while users who are generally more active on Twitter are more likely to respond, as predicted in H1a, a user who has been more active in a given conversation may actually be *less* likely to respond. We measure conversation activity (`comments count`) as the number of comments made prior to time t. While initially surprising at the user level, this finding lends support to H2b. That is, the difference in these effects may indicate that conversations have a natural ending point where users do not re-engage because they have nothing more to add, or that users may suffer from conversation fatigue—eventually getting bored or tired of engaging in the same back-and-forth.

Table 4 shows the effects for the candidate respondent's previous tweet in the conversation. Note that in addition to the features discussed below, Table 4 also shows our controls for time-varying effects using cyclic transformations of hours and days of week, which control for periodicities such as the tendency to reply more in the evenings or on weekends

Table 4. Response Predictors: Candidate Respondent's Previous Tweet.

		Significance after p correction		
	Coefficient	FDR	Cluster	FDR+Cluster
prev response	0.883	***	***	***
favorite count	–0.311	***		
retweet count	–262.234	*		
reply count	0.141	***		
quality	262.523	*		
source	0.037	***		
xday	0.169	***	**	
yday	0.239	***	*	
xhour	0.048	***		
yhour	0.193	***	*	
chars	0.367	***	***	**
has url	0.037	***		
mentions	0.155	***		
hashtags	–0.078	***	**	*
sentiment	0.362	***	*	
vader neg	0.641	***	***	**
vader pos	–0.313	***	**	*
valence	–0.084	***		
arousal	0.151	***		
dominance	–0.174	***		
time since prev	–0.658	***	***	**
topic 2	1.853	***	**	
topic 3	–0.037			
topic 4	–0.364	***		
topic 5	0.246	***		
topic 6	–0.536	***		
topic 7	–1.153	***		
topic 8	–2.787	***	***	**
topic 9	–0.573	***		
topic 10	2.404	***	**	*

Note. FDR = false discovery rate.
*$p < .1$. **$p < .05$. ***$p < .01$.

Table 5. Response Predictors: Conversation Features.

		Significance after p correction		
	Coef	FDR	Cluster	FDR + Cluster
participants	–0.179	***		
thread length	0.105	***		
thread length^2	–0.026	***	***	**

Note. FDR = false discovery rate.
*$p < .1$. **$p < .05$. ***$p < .01$.

(Cox, 2006). Many of these features are strongly predictive of response even after various error corrections. In support of H1c, users are significantly more likely to respond if they were the author of the previous response (`prev response`), that is, if we are predicting response to a tweet which was in turn a response to the candidate user. Similarly, candidates who have not engaged in the recent conversation (`time since prev`) become less likely to rejoin as time goes on. Taken together, these results may indicate that, while ostensibly a platform for multi-person conversation, dialogue on Twitter may be largely rapid-fire and dyadic in nature.

We find some evidence to suggest that candidates whose most recent tweet was longer (`chars`) are more likely to return to the conversation. This may indicate that tweet length reflects a user's enthusiasm for the conversation or for tweeting in general. Conversely, having used a hashtag in their previous tweet is negatively related to a further response, perhaps either because the purpose of the first response was to promulgate the hashtag and that job is done, or because the interlocutor did not respond in kind, demotivating the candidate respondent.

We also see interesting effects around the emotions of a candidate's most recent tweet. Based on a tweet's VADER score, it appears that, in support of H1e, if a candidates' most recent tweet was negative, they are more likely to maintain engagement in a conversation, whereas if their tweet was positive, they are less likely to continue interacting. We will return to interactive emotional dynamics in "Response Predictors: Current Tweet and Author" section.

Finally, in this corpus, users whose previous tweets focused on `Topics 8` and `10` are more likely to return to the conversation. `Topic 8` seems to indicate negative views of the democratic party, while `Topic 10` may indicate a level of nationalistic pride. This suggests that people who engage with these topics tend to be more argumentative and less likely to let a debate go than the average Twitter user in our sample.

Response Predictors: Conversation

Table 5 summarizes the effects of the features of the conversation thread itself, specifically the number of `participants` up until t and the `thread length`, as measured by the number of tweets in the conversation at time t. While at first it appears that thread length serves as a positive indicator of the interest in a thread (H2a), we do find this effect decreases for longer threads (H2b), as indicated by the quadratic `thread length`^{2} term. However, by setting the first derivative of the resulting curve to 0, we find that a reply is maximally likely for a thread length of 2—the minimum possible in our dataset. This emphasizes the tendency for nonreply and suggests that the likelihood of a thread continuing decreases monotonically as a function of thread length as users lose interest, feel the conversation is exhausted, have difficulty viewing or following the conversation, or simply move on to other things.

Response Predictors: Current Tweet and Author

Perhaps the most interesting predictors are those involving the current tweet that may or may not be responded to. This is the area in which our model extends beyond previous

Table 6. Response Predictors: Current Tweet Author.

		Significance after *p* correction		
	Coefficient	FDR	Cluster	FDR + Cluster
verified	0.069	***	*	
followers count	0.889	***		
following count	0.334	***	**	
statuses count	–0.031	***		
favourites count	–0.202	***		
comments count	0.637	***	***	**

Note. FDR = false discovery rate.
*p < .1. **p < .05. ***p < .01.

Table 7. Response Predictors: Current Tweet.

		Significance after *p* correction		
	Coefficient	FDR	Cluster	FDR + Cluster
favorite count	1.657	***		
retweet count	9.171	***		
reply count	–10.055		***	***
quality	–43.260			
source	–0.348	***	***	***
xday	–0.354	***	***	**
yday	–0.345	***	***	***
xhour	0.146	***		
yhour	–0.044	***		
chars	0.649	***	***	***
has url	0.075	***		
mentions	–0.412	***	**	
hashtags	–0.083	***	**	
sentiment	–0.135	***		
vader neg	–0.111	***		
vader pos	0.152	***		
valence	–0.524	***	**	*
arousal	–0.116	***		
dominance	0.343	***	*	
topic 2	0.815	***		
topic 3	2.140	***	**	*
topic 4	1.541	***		
topic 5	2.913	***	***	**
topic 6	1.024	***		
topic 7	2.669	***	**	*
topic 8	1.537	***		
topic 9	–0.148	***		
topic 10	2.043	***	*	
difference	–0.036	***		
Difference^2	0.188	***	**	

Note. FDR = false discovery rate.
*p < .1. **p < .05. ***p < .01.

efforts (Artzi et al., 2012; Feng & Wang, 2013; Hong et al., 2013), in that we are examining not just engagement in a conversation but responses and re-engagement at specific moments and in response to specific tweets.

The coefficient estimates for the tweet for which we are predicting replies are listed in Table 7 while the estimated coefficients for the features of that tweet's author are displayed in Table 6. For the latter, we see that, as predicted in H3a, users who are more active in the conversation, for example, have contributed more `comments`, are more likely to receive replies to their tweets. While users with verified accounts and more followers may be slightly more likely to receive replies, there is little evidence to support H3b: that the popularity of a user mediates whether or not their tweets receive a response. This may be in part because we focus only on users who have engaged at least once in the conversation, excluding the initial tweet. Thus, while an extremely popular user or entity may kick off a discussion—posting a tweet that receives numerous responses—the people who actually engage in back and forth conversations seem to be less affected by the popularity of their interlocutors.

But at best, author-level characteristics are indirect effects unless the respondent actually knows the author; the most interesting and direct effects—as well as those potentially most subject to manipulation by the tweeter—are via the tweet itself, as shown in Table 7. As before, we find that day of the week has a strong effect, though time of day has less of an effect after clustering. Interestingly, the `source` from which a tweet is posted—iPhone, Android, web interface, or third party software—also seems to be strongly predictive, with mobile users more likely to respond. This may suggest latent features of users, or simply linguistic variation of tweets mediated by the platform interface. Users posting from digital devices, for example, may be more succinct in their tweets.

In contrast to the behavior predicted by H3c, the previous `reply count` of a tweet may have a negative effect on whether a tweet will receive additional responses. That is, for each potential reply, we consider how many replies, if any, had been made up to that point. The negative effect of this coefficient suggests that there are diminishing returns to the number of replies a single tweet within a conversation thread is likely to receive, and that, reasonably enough, one may feel a tweet has been adequately rebutted if many others have already replied.

Interestingly, while tweets with more characters (`chars`) are more likely to receive replies, contra H3d, `mentions` and `hashtags` were negative indicators for our dataset. This may suggest that while longer tweets have more content to reply to, those who are overwhelmed by too many hashtags or "@"-mentions are less engaged. In addition, in support of H3e, we see that tweets with higher `valence`, for example, a higher number of "pleasant" words, are less likely to receive replies. This is similar to our previous result that if a potential respondent's previous tweet in the conversation was positive, they are less likely to reply again. Conversations about President Trump are, perhaps unsurprisingly, not just negative, but require negativity to persist.

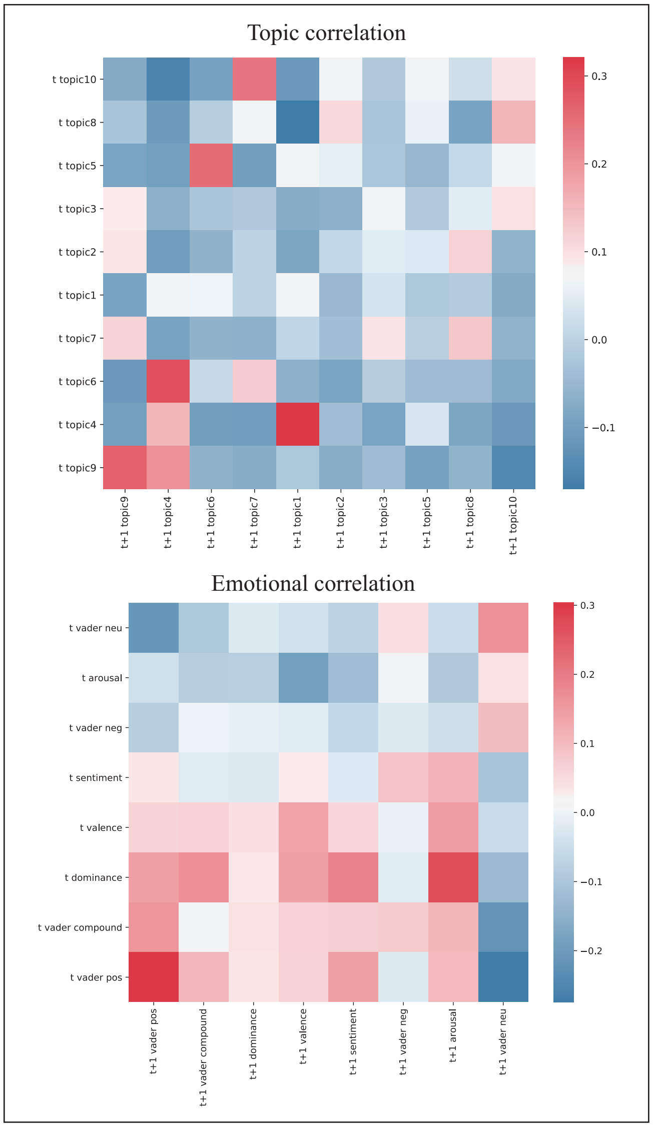

Figure 3. Correlation between tweet t and response.

Tweets focused on Topics 3, 5, and 7, were more likely to receive replies. Two of these topics, 3 and 7, are both focused on the humanitarian crises in Puerto Rico, indicating that this was a particularly active topic of back-and-forth discussion in our sample. Topic 5 is more diffuse, but points to issues of racism with words like "black," "white," and "racist" weighted within the Top 5. This topic may or may not have been tied to Puerto Rico, but indicates another area of fervent debate around the president.

Finally, the last two rows in Table 7 show the difference between the current tweet and the candidate respondent's previous tweet. This is perhaps the most psychologically interesting variable, as it speaks to a deep question about who we choose to converse with: those most like ourselves (presumably to agree), those most unlike ourselves (presumably to disagree), or something in between. We find that respondents are more likely to reply to comments very unlike their previous comment. Interestingly, the estimated minimum reply likelihood is at a distance of 0.1, which is about 1.5 standard deviations below the mean reply distance. This suggests that while there is a slight tendency to respond to comments very similar to your own (distance = 0), for most users, the likelihood of reply monotonically increases with distance. In a subsequent project, we are designing a Bayesian model of latent ideology to more precisely infer the ideological underpinnings of these sorts of political arguments.

Contents of Responses

Although this project focuses mainly on the decision to reply or not reply, rather than on the content of those replies, we can briefly examine the interactions between the content of the current tweet and the contents of its replies. These results are mainly suggestive at this point, as we do not embed this within a nested model that also controls for the decision to reply as a first stage.

Figure 3 provides heat map illustrations of correlations between tweets and their replies, both on topics (left) and emotions (right). Red indicates positive correlations, while blue shows negative correlations. Topics are sorted by the first eigenvector of the correlation matrix to cluster similar topical or emotional response patterns together.

On topics, we see that for this corpus, debate around NFL players kneeling during the national anthem seems to be highly contained, with tweets on this topic (Topic 9) often receiving responses on the same topic. On the contrary, the conversation about disaster relief following a devastating hurricane in Puerto Rico seems to be more diffuse. For example, Topic 4, which appears to largely be about the mayor of San Juan, is somewhat correlated with itself, but is more strongly correlated with Topic 1—a topic focusing on another democratic, female politician: former Secretary of State Hillary Clinton. Tweets primarily engaged in Topic 5, which seems to focus on racism, most commonly receive responses focused on Topic 6. While somewhat less coherent than other topics, it is telling that Topic 6 is most commonly followed by Topic 4. As Topic 4 praises the work of the mayor of San Juan, this may suggest our corpus contains conversations in which users are arguing about whether disaster response in Puerto Rico is tied to racism.

In addition, there are negative correlations between topics. Topic 1 and Topic 8, for example, seem to represent different views on the Democratic party. Topic 8, which potentially expresses negative views of the party, has a strong negative correlation with Topic 1, which seems to express positive views on the subject. Unsurprisingly, negative judgments of the Democratic party are rarely met with positive ones as a response, but instead some other form of negative attack.

We can also see some of these dynamics in the purely emotional content of tweets and replies. Tweets with a positive VADER score are likely to receive replies which also have a positive VADER score, even if they are less likely to receive a reply at all (as we saw earlier). Tweets scored as neutral (rather than negative) are least likely to follow positive tweets, suggesting that many of these conversations consist of like-minded people reinforcing each other's beliefs or using charged language for anyone who disagrees. Tweets which score high on the dominance measure are most highly correlated with arousal, indicating that words of strength and power are met with words of excitement—either to eagerly agree or to voraciously disagree. Neutral tweets are met with neutral tweets and are unlikely to elicit a positive response, but are also presumably less likely to receive any reply at all.

Conclusion

To summarize our results, we find that a number of user-, thread-, and tweet-level features are critical in predicting the dynamics of online conversations. While previous studies (Artzi et al., 2012; Backstrom et al., 2013; Feng & Wang, 2013; Hong et al., 2013) have primarily focused on predicting post-level engagement for the purposes of algorithmic content curation, we predict comment-level engagement as users exit and re-enter a conversation. Particularly novel is our inclusion of the features pertaining to the individual tweet that may be responded to, particularly the emotional and topical content of those tweets. Our logistic regression model predicts user response remarkably well, achieving 94% out-of-sample accuracy. The 98% accuracy achieved by our SVM model suggests that there may be further nonlinear and interactive effects among our features to explore in later work, perhaps via additional machine learning methods (such as random forests or deep neural networks); although at 98%, we are already near the ceiling of predictive accuracy.

We find support for many of the hypotheses outlined in "Features" section. In support of H1a, H1b, and H3a, we find that features of both of candidate respondents and current tweet authors have small but important effects on predicting response. Interestingly, in contrast to H3b, we find that the popularity of a tweet's author has little effect on predicting whether or not a tweet will receive a reply. Because we only include users who have been active in a back-and-forth exchange, this suggests that conversations on Twitter are relatively free from the sort of social influence we may have found if we were examining replies to the initial tweet of a thread.

In support of H2b, we find that longer conversations are decreasingly likely to receive replies. Similarly, while in H1c we expected a candidate's previous activity in a thread to predict additional activity, we found that users who have already offered numerous comments are less likely to re-engage. Taken together, this suggests that the time attention and cognitive energy that goes into participating in a back-and-forth exchange lead to a natural cutoff where conversations, though popular at first, become too much effort to continue or have their subject matter exhausted.

In addition, and in support of H1d and H3e, we find that the emotional and topical contents of tweets seem to play a significant role in driving the continuation of conversation. Users with negative-sentiment tweets are more likely to re-enter conversations, and tweets with fewer pleasant words are more likely to receive a response. While a small corner of our corpus may be primarily engaged in positive-to-positive conversations, it seems that the vast majority of Twitter dialogue around President Trump consists of acrimonious argumentation. This is reinforced by the finding that candidates are more likely to respond to tweets *unlike* their previous tweet. This could optimistically be interpreted as people engaging in dialogue across difference, but could just as well be mutual trolling—though if the latter, at least we do observe extensive repeated interactions rather than simple one-off attacks.

These findings paint a picture consistent with what some avid Twitter users might expect: At least when it comes to political dialogue around a controversial figure, most conversations are emotionally charged and negative in tenor. While such conversations fall far below the ideals of democratic deliberation (Cohen, 1989; Dewey, 1927; Habermas, 1984), our findings suggest this may not be the end of the story. First, within the broader deliberative system, it is commonly acknowledged that many "everyday" conversations will frequently fail to meet deliberative ideals (Mansbridge, 1999). Nevertheless, these conversations may still play an important and positive role in expanding people's view points and encouraging refinement of beliefs (Huckfeldt et al., 2004; Mansbridge, 1999). Despite the predominately negative tone of our corpus, we do find that users are responding to topics outside their own talking points, and that users who are active in a conversation are more likely to receive a response. In other words, conversations *are* happening, and those conversations do not appear to be strictly confined within partisan bubbles.

Furthermore, it is difficult to fully characterize the democratic value of a conversation based on sentiment analysis alone. For example, within our corpus, we find that positive tweets are most likely to receive positive-sentiment responses. While, on the surface, this may suggest a collection of more civil exchanges, it is also possible that positive-to-positive conversations represent little more than in-party affirmation, with little deliberative value. The deliberative ideal imagines citizens as "tolerant gladiators" (Huckfeldt et al., 2004), who fight with strong words but who emerge from confrontation as friends. Our corpus finds evidence that there is no shortage of strong words—rather, its the long-term effects of these conversations which remain to be seen.

Taken together, these findings suggest that average citizens are participating in rich and engaging political conversations. While the extent to which these conversations support democratic ideals remains to be seen, if we wish to extend and enrich these interactions, we should seek to broadly increase conversational activity online, developing tools to make it easier to engage and follow long threads. In addition, given user's willingness to respond to those unlike themselves, these findings suggest that there is value in adding noise to recommender systems—showing users new and different content, rather than overfitting recommendations based on the content with which they have already interacted.

In future work, we intend to analyze the dynamics of political conversations over time, looking at sentiment and opinion flows through whole threads of conversation, and we are developing a model to infer the role of latent ideology in these exchanges. In this article, we have identified a number of key factors predictive of conversation engagement and shown that these conversations are by no means chaotic, but in fact are systematic and highly predictable, reflecting a complex interplay of circumstance, topic, and emotion.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. While there are a broad range of users/entities with "verified" status, the badge is intended to indicate authentic accounts "of public interest" (https://help.twitter.com/en/managing-your-account/twitter-verified-accounts).
2. Following the Twitter API's naming conventions, we use "`favorite count`" as the tweet-level feature measuring the number of times a tweet as been liked and "`favourites count`" as the user-level feature indicating the number of tweets a user has liked in their account lifetime.
3. This number of topics was selected because it yields meaningful topics which are coherent to a human reader, but is relatively arbitrary; topic counts of 5, 15, or 20 all produce similar results.
4. Using an 80% in-sample, 20% out-sample split.
5. We use the Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini & Hochberg, 1995), setting a false positive rate of 10%.

References

Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the 17th international conference on World Wide Web* (pp. 665-674). New York, NY: Association for Computing Machinery.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, KDD '12* (pp. 850-858). New York, NY: Association for Computing Machinery.

Artzi, Y., Pantel, P., & Gamon, M. (2012). Predicting responses to microblog posts. In *Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12* (pp. 602-606). Stroudsburg, PA: Association for Computational Linguistics.

Axelrod, R. (1987). The evolution of strategies in the iterated prisoners dilemma. In L. Davis (Ed.), *Genetic algorithms and simulated annealing* (pp. 32-41). London: Pitman, and Los Altos, CA: Morgan Kaufman.

Backstrom, L., Kleinberg, J., Lee, L., & Danescu-Niculescu-Mizil, C. (2013). Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proceedings of the sixth ACM international conference on Web Search and Data Mining, WSDM '13* (pp. 13-22). New York, NY: Association for Computing Machinery.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the fourth ACM international conference on Web Search and Data Mining, WSDM '11* (pp. 65-74). New York, NY: Association for Computing Machinery.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*, 1130-1132.

Bednar, J., & Page, S. (2007). Can game (s) theory explain culture? The emergence of cultural behavior within multiple games. *Rationality and Society*, *19*, 65-97.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*, 289-300.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144-152). New York, NY: Association for Computing Machinery.

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., & Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 661-670). Association for Computing Machinery.

Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J., & Bernstein, M. (2017). Anyone can become a troll: causes of trolling behavior in online siscussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1217-1230). New York, NY: Association for Computing Machinery.

Cohen, J. (1989). Deliberation and democratic legitimacy. In A. P. Hamlin & P. Pettit (Eds.), *The good polity: Normative analysis of the state* (pp. 17-34). New York, NY: Blackwell.

Cox, N. J. (2006). Speaking stata: In praise of trigonometric predictors. *The Stata Journal*, *6*, 561-579.

Dewey, J. (1927). *The public and its problems*. Athens, OH: Swallow Press.

Farina, C., Epstein, D. B., Heidt, J. J., & Newhart, M. (2013). Regulation room: Getting "more, better" civic participation in complex government policymaking. *Transforming Government: People, Process and Policy*, *7*, 501-516.

Feng, W., & Wang, J. (2013). Retweet or not? Personalized tweet re-ranking. In Proceedings of the sixth ACM international conference on Web Search and Data Mining, WSDM '13 (pp. 577-586). New York, NY: Association for Computing Machinery.

Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 101-110).

George, F., & Kibria, B. G. (2011). Confidence intervals for signal to noise ratio of a Poisson distribution. *American Journal of Biostatistics*, *2*(2), 44-55.

Habermas, J. (1984). *The theory of communicative action*. Boston, MA: Beacon Press.

He, Y., & Tan, J. (2015). Study on sina micro-blog personalized recommendation based on semantic network. *Expert Systems With Applications*, *42*, 4797-4804.

Hong, L., Doumith, A. S., & Davison, B. D. (2013). Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on Wweb Ssearch and Ddata Mmining, WSDM '13* (pp. 557-566). New York, NY: Association for Computing Machinery.

Huckfeldt, R., Mendez, J. M., & Osborn, T. (2004). Disagreement, ambivalence, and engagement: The political consequences of heterogeneous networks. *Political Psychology*, *25*, 65-95.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the eighth international AAAI conference on weblogs and social media* (pp. 216-225). Palo Alto, CA: AAAI Press.

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., . . . Xie, L. (2012). Social media use by government: From the routine to the critical. *Government Information Quarterly*, *29*, 480-491.

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, *28*, 331-339.

Lo, A., Chernoff, H., Zheng, T., & Lo, S. H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 13892-13897.

Mansbridge, J. (1999). Everyday talk in the deliberative system. In S. Macedo (Ed.), *Deliberative politics: Essays on democracy and disagreement* (pp. 1-211). New York, NY: Oxford University Press.

Mansbridge, J. (2015). A minimalist definition of deliberation. In P. Heller & V. Rao (Eds.), *Equity and Development Series: Deliberation and development: Rethinking the Role of voice and collective action in unequal societies* (Vol. 2, pp. 27-50). Washington, DC: World Bank.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*, 629-649.

Myers, S. A., & Leskovec, J. (2014). The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World Wide Web, WWW '14* (pp. 913-924). New York, NY: Association for Computing Machinery.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A. S. Dadzie, & M. Hardey (Eds.), *Proceedings of the ESWC2011 workshop on "making sense of microposts": Big things come in small packages, CEUR workshop proceedings* (pp. 93-98). Heraklion, Greece: European Semantic Web Conference.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *International Conference on Weblogs and Social Media*, *11*, 122-129.

Oktay, H., Taylor, B. J., & Jensen, D. D. (2010). Causal discovery in social media using quasi-experimental designs. In *Proceedings of the first workshop on social media analytics* (pp. 1-9). New York, NY: Association for Computing Machinery.

Pan, Y., Cong, F., Chen, K., & Yu, Y. (2013). Diffusion-aware personalized social update recommendation. In *Proceedings of the 7th ACM conference on recommender systems* (pp. 69-76). New York, NY: Association for Computing Machinery.

Sanders, L. M. (1997). Against deliberation. *Political Theory*, *25*, 347-376.

Shen, H. W., Wang, D., Song, C., & Barabási, A. L. (2014). Modeling and predicting popularity dynamics via reinforced Poisson processes. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence* (Vol. 14, pp. 291-297). Palo Alto, CA: AAAI Press.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289-310.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, *10*, 175-195.

Vosecky, J., Leung, K. W. T., & Ng, W. (2014). Collaborative personalized twitter search with topic-language models. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 53-62). New York, NY: Association for Computing Machinery.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191-1207.

Yan, R., Lapata, M., & Li, X. (2012). Tweet recommendation with graph co-ranking. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers* (Vol. 1, pp. 516-525). New York, NY: Association for Computational Linguistics.

Author Biographies

Sarah Shugars is a doctoral candidate at the Network Science Institute, Northeastern University. Her research focuses on American political behavior, public opinion, and deliberative democracy and engages a range of methods including machine learning, natural language processing, and network science.

Nicholas Beauchamp is an assistant professor in the Department of Political Science, Northeasten University. His research examines how political opinions form and change as a result of discussion, deliberation and argument in political domains using techniques from machine learning, automated text analysis, and social network analysis.