Violent Political Rhetoric on Twitter *

TAEGYOON KIM

Violent hostility between ordinary partisans is undermining American democracy. Social media is blamed for rhetoric threatening violence against political opponents and implicated in offline political violence. Focusing on Twitter, I propose a method to identify such rhetoric and investigate substantive patterns associated with it. Using a data set surrounding the 2020 Presidential Election, I demonstrate that violent tweets closely track contentious politics offline, peaking in the days preceding the Capitol Riot. Women and Republican politicians are targeted with such tweets more frequently than non-Republican and men politicians. Violent tweets, while rare, spread widely through communication networks, reaching those without direct ties to violent users on the fringe of the networks. This paper is the first to make sense of violent partisan hostility expressed online, contributing to the fields of partisanship, contentious politics, and political communication.

The emergence of social media platforms was widely touted as a technological revolution that would bring about many beneficial outcomes such as political learning and participation (Dimitrova et al. 2014; Tucker et al. 2017). However, such early hopes are being overshadowed by mounting concerns about aggressive political communication. In recent days, one can easily encounter uncivil political discussion

*Taegyoon Kim is a dual-title Ph.D. candidate in Political Science and Social Data Analytics, Pennsylvania State University (taegyoon@psu.edu).

both from political elites as well as ordinary users. Also, various types of hate speech — targeted at women, ethnic minorities, and partisan opponents — are common and viral on social media (Mathew et al. 2019). Accordingly, much scholarly attention has been paid to detect such speech and curb its spread (Siegel 2020). However, we know very little about another, perhaps most deleterious, type of aggressive political speech: violent political rhetoric. Violent political rhetoric, expressing the intention of physical harm against political opponents, has drawn significant media attention. Numerous media reports show that malevolent users on social media write posts that threaten violence against political opponents on the basis of partisanship, ideology, and gender and that such posts are even associated with the actual incidences of offline violence (Brice-Saddler 2019; Daugherty 2019; Vigdor 2019). In particular, many social media platforms are implicated in the extremist effort to motivate and organize the Capitol Riot that left a vivid and deep scar on American democracy. Plenty of evidence shows that not only niche extremist online forums but also mainstream social media platforms, including Twitter, were exploited by users who called for violence in the days preceding the riot on January 6 2021 (Guynn 2021; Lytvynenko and Hensley-Clancy 2021; Romm 2021).

Violent political rhetoric is worrisome not only because it serves as a harbinger of extremist offline violence but also because exposure to such rhetoric has harmful consequences such as increased tolerance for offline violence against political opponents (Kalmoe 2014) and ideological polarization (Kalmoe, Gubler, and Wood 2018). It is particularly concerning because violent political rhetoric can widely spread through the communication network on social media, amplifying its negative effects. Besides, such rhetoric is in itself a behavioral manifestation of violent partisanship where individuals not just hate out-partisans (Abramowitz and S. W. Webster 2018) but also support and even enjoy the use of violence against them (Kalmoe and Mason 2018). The rhetoric is an

online mirror image of the recent instances of inter-partisan offline violence surrounding contentious political issues (e.g., Black Lives Matter movements, the controversies about the 2020 Presidential Election) and is no less concerning than its offline counterpart (Pilkington and Levine 2020).

How prevalent is violent political rhetoric on social media? How do posts containing such rhetoric relate to offline-world politics? What types of politicians are targeted? What users use violent rhetoric against political opponents? How diffusive is violent political rhetoric and what predicts its spread? Given the significance of violent political rhetoric, it is urgent to investigate these questions. Due to the massive size of the content generated in real-time, however, it is prohibitively expensive to manually identify violent content on a large scale, leaving only anecdotal and incomprehensive evidence (Lytvynenko and Hensley-Clancy 2021; Romm 2021). Therefore, I propose an automated method for detecting violent political rhetoric from a continuous stream of social media data, focused on Twitter. I then apply the method to build a data set of tweets containing violent political rhetoric over a 16-week period surrounding the 2020 Presidential Election. Finally, I provide comprehensive data analyses on the characteristics and spread of violent political rhetoric.

By doing so, I contribute to three areas of research in political science. First, I shed light on the literature on political violence by extending the study of individuals' engagement in political violence to online domains. While a body of research in offline political violence has taken a bottom-up approach to study individuals who take part in collective violence in the offline world (Claassen 2016; Fujii 2011; Horowitz 1985; Scacco 2010; Tausch et al. 2011), few studies have taken a similar approach to investigate individuals who threaten violence against political opponents in online space. I fill part of the gap by showing that individuals who threaten violence against political opponents on social

media are ideologically extreme and located on the fringe of the online communication network. I also show that they threaten opposition politicians, in the context of heightened contentious politics offline. The online-offline links identified in my study open up a future research agenda on what causal mechanisms connect threats of political violence online and contentious offline politics (including offline political violence).

By identifying and characterizing violent political rhetoric on Twitter, I also extend the study of aggressive online political communication where incivility and hate speech have been the key areas of inquiry (Berry and Sobieraj 2013; Gervais 2015, 2019; Munger 2017, 2021; Popan et al. 2019; Siegel 2020; Siegel et al. 2021; Suhay, Bello-Pardo, and Maurer 2018; Sydnor 2019). Building on a new data set spanning the crucial period surrounding the 2020 Presidential Election, I show that, although tweets containing violent political rhetoric are rare (0.07% of political tweets, on average), they spread beyond direct ties to violent users. I find that almost 40% of the retweets of such content spread through indirect ties (i.e., my friend's friend, a friend of my friend's friend, etc.), thereby creating huge potential for incidental exposure to such abhorrent language. I also demonstrate that, although threatening tweets are shared primarily among ideologically similar users, there is a considerable amount of cross-ideological exposure as well, calling for further investigation into the effects of exposure to violent political rhetoric both from an in-party member and from an out-party member.

Finally, I shed light on the literature on mass partisan polarization and negative partisanship by demonstrating that violent partisanship is manifested online in the form of threats against partisan opponents. Recent studies on mass partisan polarization highlight that partisans are not just ideologically far apart (Abramowitz and Saunders 2008; Fiorina and Abrams 2008) but also dislike or even endorse violence against our-party members (Abramowitz and S. W. Webster 2018; Iyengar, Sood, and Lelkes 2012; Iyengar et al.

2019; Kalmoe and Mason 2018). However, there was little effort to explore how violent partisanship is expressed online. My work contributes to the literature by providing an easy-to-access indicator for tracking the level of violent partisanship. Considering the evidence that there are significant discrepancies between survey self-reports and actual online behavior (Guess et al. 2019), my study provides an excellent complement to survey-based measurement as it enables researchers to directly observe the over-time trend of violent partisan behavior expressed online.[1] For instance, I illustrate that the level of violent political rhetoric on Twitter corresponds to the violent partisan tension offline, reaching its peak in the days preceding the Capitol Riot.

Related Work

In this paper, I build on and contribute to three streams of literature. First, a large body of works takes a micro-level approach to individuals' participation in offline political violence, providing rich theoretical insight into individuals threatening political opponents online. Second, an extensive body of research in political communication investigates the consequences of politicians' use of violent political metaphors in the offline world and explores aggressive speech in online political discussion, shedding light on a rich context for an inquiry into violent political rhetoric online. Third, violent political rhetoric on social media is a form of behavioral manifestation of extreme negative partisanship. The literature on political polarization and negative partisanship is very useful in understanding why social media users express a violent intention against out-partisans and what consequences

[1]Although this approach shares with survey self-reports a concern that they both can be susceptible to intentional exaggeration or suppression resulting from social norms, the former nonetheless is far less reactive than the latter.

such behavior has.

Offline Political Violence

Although few studies exist to explain political violence online, there is an extensive body of literature devoted to explaining why individuals engage in offline political violence in various settings. Focused on conflict-ridden contexts, studies seek to explain why individuals participate in inter-group violence (ethnic, religious, partisan). Major explanations include selective incentives that enable individuals to overcome the problem of free-riding (DiPasquale and Glaeser 1998; Humphreys and Weinstein 2008), social pressure (Fujii 2011; Scacco 2010), and perceived distributive inequality (Claassen 2016). Also, an interdisciplinary stream of studies on violent extremism seeks to identify a host of risk factors associated with individuals' tendency to join violent extremist activities (Borum 2011a, 2011b; Gill, Horgan, and Deckert 2014; LaFree and Ackerman 2009; McGilloway, Ghosh, and Bhui 2015). Lack of stable employment, history of mental illness, low self-control, perceived injustice, and exposure to violent extremism are among the factors highlighted in the literature (LaFree et al. 2018; Pauwels and Heylen 2017; Schils and Pauwels 2016).

Aggressive Political Communication

Raising concerns about political elites' violent rhetoric in the U.S., a recent strand of studies investigates its political consequences (Kalmoe 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019; Matsumoto, Frank, and Hwang 2015). Kalmoe (2019) shows that violent political metaphors (metaphors that describe politics as violent events such as a battle or a war) increase willingness to vote among individuals with highly aggressive personalities but

the opposite effect is found among individuals low in aggressive personalities. Focusing on issue polarization, Kalmoe, Gubler, and Wood (2018) finds that violent political metaphors prime aggression in aggressive partisans and thus lead to intransigence on issue positions.

While violent political rhetoric is mainly studied in context of political elites' offline speech, many works in online political communication focus on incivility and hate speech. They point out that the reduced gate-keeping power of traditional media outlets and online anonymity gave rise to uncivil and hateful content targeted at people of a different race, gender, and partisan affiliation (Berry and Sobieraj 2013; Kennedy and Taylor 2010; Munger 2017, 2021; Shandwick 2019). Aggressive online speech is reported to have crucial consequences for a host of political behavior, including participation (Henson, Reyns, and Fisher 2013; Sydnor 2019), information seeking (Sydnor 2019), inter-group evaluations, and deliberative attitudes (Gervais 2019).[2] Accordingly, a large body of works are devoted to detecting (Davidson et al. 2017; Siegel 2020; Waseem and Hovy 2016; Zimmerman, Kruschwitz, and Fox 2018) and discouraging hateful speech (Munger 2017, 2021).

Affective Polarization and Negative Partisanship

Recent scholarship on political polarization highlights affective polarization, the degree to which citizens dislike and distrust out-partisans (Iyengar et al. 2019; Iyengar, Sood, and Lelkes 2012). Pointing out the increasing affective polarization over the last several decades (Iyengar et al. 2019), the scholarship seeks to uncover its inimical consequences, including anti-deliberative attitudes, social avoidance, and outright social discrimination

[2]For a comprehensive review of behavioral consequences of political incivility and a discussion of related psychological processes, see Sydnor 2019.

(Abramowitz and S. Webster 2016; Broockman, Kalla, and Westwood 2020; Druckman et al. 2020; Huber and Malhotra 2017; Hutchens, Hmielowski, and Beam 2019; Iyengar, Sood, and Lelkes 2012; MacKuen et al. 2010). Extending the study of negative partisanship, some works take one step further, evaluating the extent to which partisans rationalize harm and even endorse violence against partisan opponents (Kalmoe and Mason 2018; Westwood et al. 2021). Such negative partisanship has mainly been measured using survey self-reports. While there exist a handful of other approaches, such as IAT (Implicit Association Test) (Iyengar et al. 2019), survey self-reports have been the only strategy to measuring violent partisanship (Kalmoe and Mason 2018; Westwood et al. 2021).

Targeted Violent Political Rhetoric

Building on the psychology literature on aggression (Anderson and Bushman 2002), I define violent political rhetoric as rhetoric expressing the intention of severe physical harm against political opponents.[3] This involves a threat and support of physical harm against political opponents and hopes of extreme physical harm inflicted on them (*schadenfreude*).[4] Here, violent political rhetoric is targeted at a specific political entity. The target is typically a partisan opponent, either a group (e.g., Republican representatives, Democratic senators) or an individual politician (e.g., Donald Trump, Joe Biden).

Existing studies on violent political rhetoric have employed various conceptualizations (Kalmoe 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019; Zeitzoff 2020). Zeitzoff

[3]Anderson and Bushman (2002) defines aggression as "any behavior directed toward another individual that is carried out with the proximate (immediate) intent to cause harm" and violence as a form of "aggression that has extreme harm as its goal."

[4]There is an active discussion on how to conceptualize and measure political violence in survey research (Kalmoe and Mason 2018; Westwood et al. 2021).

(2020) employs an expansive definition of violent political rhetoric: "any type of language that defames, dehumanizes, is derogatory, or threatens opponents." Thus, violent political rhetoric is conceptualized as a spectrum that encompasses "name-calling and incivility at the lower end and threats or calls for violence at the upper end." Closely related to my study is the type of violent political rhetoric at the upper end of the spectrum.

Kalmoe and his coauthors' works focus specifically on violent political metaphors (Kalmoe 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019). In their work, violent political metaphors are defined as "figures of speech that cast nonviolent politics of campaigning and governing in violent terms, that portray leaders or groups as combatants, that depict political objects as weapons, or that describe political environments as sites of non-literal violence." In contrast to the definition employed in my study, this type of violent political rhetoric does not threaten (or support, incite) any physical violence against political opponents.

Detecting Violent Political Rhetoric on Twitter

Many approaches have been proposed to detect hostile speech on social media, including incivility (Davidson, Sun, and Wojcieszak 2020; Theocharis et al. 2020) and hate speech (Siegel 2020), employing a variety of approaches from dictionary (Dadvar et al. 2012; Isbister et al. 2018; Magu, Joshi, and Luo 2017) to machine learning methods (Nikolov and Radivchev 2019; Williams et al. 2020).[5] However, there has been little effort to identify violent political rhetoric, a distinct form of hostile speech. While a small body of research

[5]Machine learning methods typically outperform dictionary methods. For a comprehensive review of works focused on detecting incivility and hate speech, see Davidson, Sun, and Wojcieszak (2020) and Siegel (2020), respectively.

on YouTube proposes several methods to identify threatening comments from YouTube videos (Hammer et al. 2019; Wester 2016; Wester et al. 2016), they are narrowly focused on a small sample of videos in a highly specific context.[6] In this section, I introduce a new method that combines keyword filtering and machine learning to detect violent political rhetoric from a massive stream of content on Twitter (Figure 1).

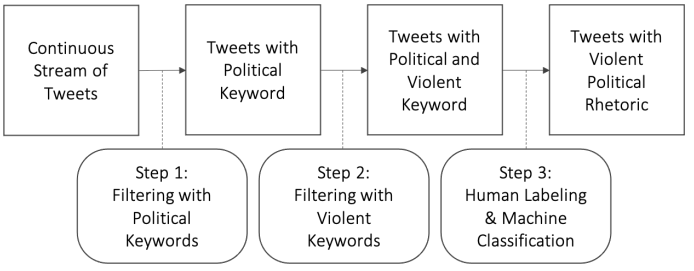


Figure 1. Data collection pipeline

Step 1: Filtering through Political Keywords

I start with compiling a list of political keywords to download tweets from the Twitter API (Application Programming Interface).[7] Since a massive number of heterogeneous tweets are generated in real-time, I first filter the tweet stream through a set of political keywords. The keywords involve a broad sample of politicians' accounts (members of Congress, governors, and the four candidates of the 2020 Presidential Election) as well as those belonging to major parties (link to the list). The tweets filtered and downloaded through the list of accounts "mention" (Twitter 2021a) at least one of the political accounts in the list.[8] Naturally, the keywords of my choice make the downloaded tweets political in

[6]See Appendix D for more information.
[7]The Twitter API is used to retrieve data and engage with the communication on Twitter.
[8]Mentions appear in tweets a) when users reply to other users' tweets (then, the account of the original tweeter automatically appears in the reply) and b) when users simply include the account in

nature. In addition, focusing on tweets mentioning these accounts is an effective approach to gather political tweets that engage (and threaten) politicians in conversation.[9]

I run a Python program that scrapes live tweets that contain any of the keywords in the list. The program is designed to scrape live tweets continuously via the Streaming API (Twitter 2021c). This API allows researchers to scrape live tweets as they are published while another major API, the Search API, provides access to historical tweets up to a certain number of days in the past (Twitter 2021f). The decision to opt out of the Search API is due to the potential for the platform to engage in censorship. That is, a set of tweets retrieved via the Search API will leave out violent tweets that have been deleted by Twitter for violating its terms of service.[10]

Step 2: Filtering through Violent Keywords

Once I have collected a corpus of tweets with at least one political keyword, I move on to the task of splitting it into violent and non-violent. Here, my approach is very similar to the one taken in the previous subsection. I first compile a list of violent keywords and filter the

their tweet text. The function is the key communicative component on Twitter with which users initiate and keep engaging with each other (Twitter 2021a).

[9]Though I do not intend to build a sample of "all political tweets", focusing on mention tweets might not represent political tweets engaging politicians in conversation. This is primarily because users still can and do reference politicians using their name ("Donald Trump" as opposed to "@realDonaldTrump"). To evaluate the extent to which focusing on mention tweets bias any downstream analysis, I calculated the proportion of the number of tweets including a given politician's full names to the number of tweets including their accounts and compared the proportion across major politician-level attributes highlighted in the analysis. I report the results in Appendix A. I find no evidence for any tendency that politicians are referenced differently in terms of the choice of the full name and the handle, across gender, political party, and position.

[10]Twitter has detailed policies on violent threats (Twitter 2021h). Essentially, its approach is post hoc in that it reviews what is already publicly published and decides whether to moderate content or sanction users. The data set I gather through the Streaming API (Twitter 2021c) avoids such post hoc moderation. In addition, to the best of my knowledge, it is unclear whether Twitter has a mechanism that prevents users from writing violent content in the first place.

existing tweets through those keywords. A challenge here is that any human-generated list of keywords might leave out potentially relevant tweets. As King, Lam, and Roberts (2017) demonstrates, humans are not particularly capable of coming up with a representative list of keywords for a certain topic or concept. In other words, it is hard for any single researcher to come up with a comprehensive set of keywords used to express a violent intention against partisan opponents (e.g., kill, shoot, choke, etc.).

To deal with this, I combine model-based extraction of keywords with human judgment. First, I start with fitting a model to score terms in an external corpus that was already human-labeled in terms of whether a text is threatening or not. Here, I intend to extract violent keywords from a corpus that already contains information about what multiple people deem to be threatening. Specifically, I use a data set built by Jigsaw, a unit within Google (Jigsaw 2020) (link to the data set). The data set contains around two-million Wikipedia comments labeled by human coders for various toxic conversational attributes, including "threat." I fit a logistic regression model and extract terms (uni- and bi-gram features) that are most predictive of perceived threat (in terms of the size of the weights assigned to them). Second, given the weighted terms, I then use human judgment to set a threshold above which terms are included in the list of violent keywords. I set the threshold at the top-200 because over the top-200 terms, the terms were too generic to indicate any intention of violence. Using the list of terms, I divided the political tweets from Step 1 into ones with and without at least one violent keyword. For more detailed information about keyword filtering in general and violent keywords, see Appendix B.

Step 3: Manual Labeling and Machine Classification

Although the previous round of filtering relies on a list of violent keywords that people frequently use online as well as consider to be violent, only a small fraction of the violent-keyword tweets contain the intention of violence. This is because many tweets contain a violent keyword without expressing any intention of physical harm against political opponents. The major sources of false positives involve a) when violent keywords are used as a metaphor that describes non-violent political events (Kalmoe 2013, 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019), b) a religious curse that does not threaten physical harm (e.g., 'burn in hell!'), c) quoting (or even criticizing) violent political rhetoric from someone else, and d) irony (e.g., 'why don't just shoot them all if you believe violence solves the problem?'). To more accurately identify tweets containing violent political rhetoric, three human coders, including myself and two undergraduate assistants, classified tweets in terms of whether the author expresses the intention of severe physical harm against a political opponent (link to the detailed coding rules). The coders manually labeled a set of 2,500 tweets together and then individually labeled over 7,500 tweets. The inter-coder agreement score in terms of Krippendorff's Alpha is around 0.6, higher than the standard in the relevant literature (Krippendorff 2018). For more information on the manual labeling, see Appendix C.

In addition, I used active learning (Linder 2017; Miller, Linder, and Mebane 2020; Settles 2009). Since the corpus compiled through Step 1 and 2 is highly imbalanced with only a small fraction containing violent political rhetoric, randomly sampling a training set for regular supervised learning will lead to inefficiency. That is, the training set will contain too few relevant tweets for any classifier to learn about what features predict violent political rhetoric. In active learning, I go through an iterative process where I start

with manually labeling *randomly sampled* texts to train a classifier, *select (not randomly)* texts whose predicted probabilities are around the decision threshold (ones whose class the classifier is most uncertain about), manually label the around-the-threshold texts, and finally accumulate those texts to re-train the classifier.

Through the iterative process, I compiled a training set of violent-keyword tweets labeled for violent rhetoric. I then trained various machine learning classifiers and the performance of the classifiers was evaluated on unseen (or held-out) data using 5-fold cross validation in terms of precision, recall, and F-1 (Han, Pei, and Kamber 2011). To label the rest of the tweets, I selected the best performing classifier (precision: 73.11, recall: 62.81, F-1 67.48), one built on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018). For more information on the active learning and machine classification process, see Appendix D.

Characteristics and Spread of Tweets Containing Violent Political Rhetoric

How prevalent is violent political rhetoric on social media? How do posts containing such rhetoric relate to offline-world politics? What types of politicians are targeted? What users use violent rhetoric against political opponents? How diffusive is violent political rhetoric and what predicts its spread? In this section, I provide comprehensive data analyses concerning the characteristics and the spread of tweets containing violent political rhetoric. The following analysis is based on a data set of tweets collected between September 23 2020 and January 8 2021.[11] This 16-week period covers major political events concerning

[11]The data set includes 343,432,844 political tweets (235,019 among them are classified as violent). For computational efficiency, I randomly sampled 1/2000 of the non-violent political tweets

the 2020 Presidential Election, including the Capitol Riot and the suspension of Trump's Twitter account.

The key findings involve the following. Violent political rhetoric on Twitter is closely related to offline contentious politics, spiking to its highest level in the days preceding the Capitol Riot. In terms of targeting, women and Republican politicians are more frequently targeted than men and non-Republican politicians. Violent users are ideologically extreme, located on the fringe of the communication network, and their ideological makeup varies over time depending on what issues violent political rhetoric arises from. The spread of violent tweets takes place primarily among ideologically similar users but there is also a substantial amount of cross-ideological spread, raising concerns about co-radicalization. While violent political rhetoric is rare (0.07% of political tweets) but almost 40% of retweets of violent tweets take place between users without a direct following tie, incidentally exposing a potentially huge audience to such appalling content.

Content and Timeline of Violent Tweets

To shed light on how tweets containing violent political rhetoric differ from non-violent political tweets in terms of content, Figure 2 shows the terms that divide non-violent political tweets from violent political tweets. I rely on a feature selection/weighting method for comparing word usage across different groups called Fightin' Words (Monroe, Colaresi, and Quinn 2008).[12] In the figure, the x-axis indicates the relative frequency with which the keyword occurs in each type of tweets. The y-axis in each panel depicts the

from each day and used the sampled tweets in the analysis.

[12]The method models word usage differences across different groups in a way that reduces the prominence of words used too frequently or too infrequently. The method produces a z-score that quantifies the significance with which the use of a word differs between the two groups of documents.

extent to which the keyword is associated with each type of tweets (see Appendix E for the top-30 keywords by type-specificity). Note that some of the words included as indicating violent political tweets have already been baked in as part of the violent-keyword filtering.

What is most noteworthy is that words that indicate certain political entities are much more frequent for violent tweets than for non-violent ones. We can see that, while no entity-specific words were included in the keywords for non-violent tweets, the violent keywords include many accounts that belong to high-profile political figures such as @realdonaldtrump (Donald Trump), @senatemajldr (Mitch McConnell), @mike_pence (Mike Pence), and @secpompeo (Mike Pompeo). In particular, the account for Trump, "@realdonaldtrump", demonstrates that he was at the center of violent and divisive communication on Twitter. The prevalence of entity-specific words is also consistent with our focus on targeted violent political rhetoric. For the words indicating non-violent tweets, many general political terms are included (e.g., presid, vote, tax, elector, campaign) along with words that represent particular political events such as "georgia" (the Senate election in Georgia) or "fraud" (misinformation about election fraud).

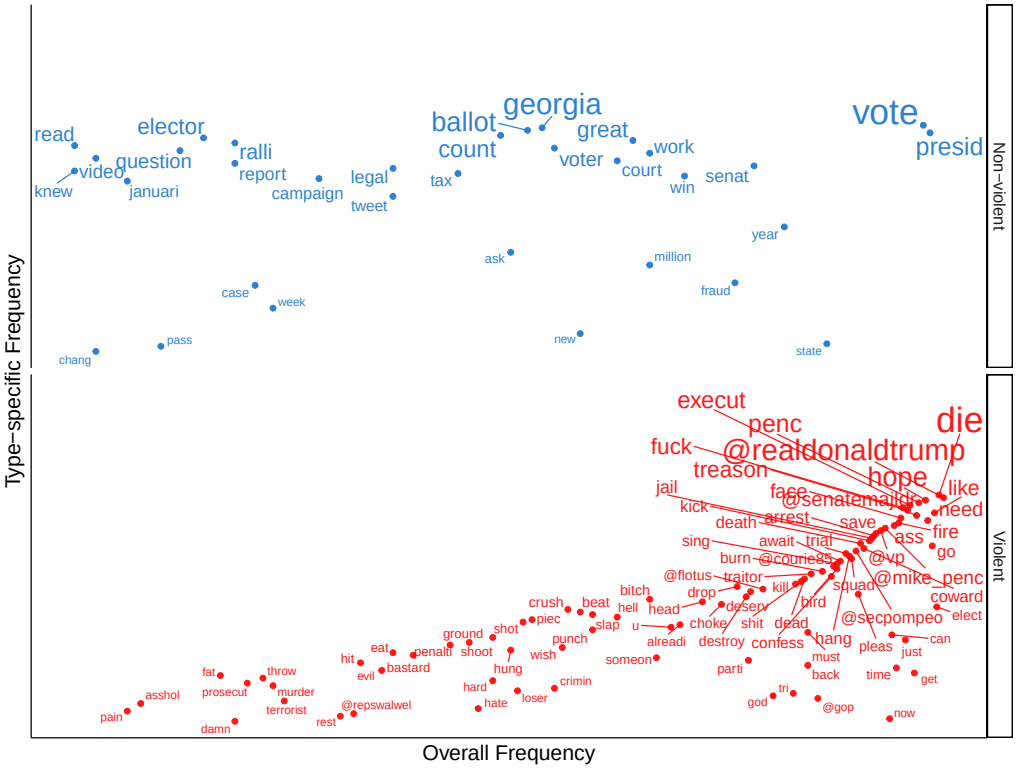


Figure 2. Comparison of terms by type of tweets

Note: For the analysis, I took a sample of 10,000 tweets, with 5,000 from each type. I used an R package `quanteda` for text preprocessing. Punctuation, symbols, numbers, stopwords, and URLs were removed from the text. The text was lower-cased and stemmed.

Now that we understand the stylistic characteristics of violent political rhetoric, what is talked about in violent tweets? To provide a general sense of the content in violent tweets,

Table 1 reports the top-30 hashtags that are most frequently used in violent tweets.[13] Note that I had lower-cased the text of the tweets before extracting hashtags to match ones that only differ in capitalization. In general, the hashtags together show that the content of violent political rhetoric is highly variegated, revolving around diverse political/social issues: general partisan hostility (#wethepeople, #1), racial conflict (#antifaarefascists, #blmareracists), moral issues (#brandonbernard, #pardonsnowden, #freeassange), election campaigning (#vote, #trump2020), disputes over the election result (#pencecard, #fightback, #1776again), and the COVID-19 pandemic (#covid19, #walterreed, #covidiot). For the hashtags reflecting general partisan hostility ("#wethepeople" and "#1"), close manual reading reveals that they are used when users emphasize their in-partisans as representing the whole country (the former) and their out-partisans as the foremost enemy of the country (the latter). Although it is beyond the scope of this study to review every hashtag in the list (some will be discussed in the next section), they together make it clear that violent political rhetoric is closely related to various political/social issues in offline politics.

[13]Not all violent tweets contain a hashtag so the partisan source of the hashtags in Table 1 (or Table 2) does not necessarily correspond to the distribution of violent users' ideology or their partisanship in the entire data set.

TABLE 1 *Most frequent hashtags in violent political rhetoric (entire period)*

Rank	Hashtag	Count	Rank	Hashtag	Count
1	#wethepeople	1,511	16	#pardonsnowden	365
2	#1	1,398	17	#traitortrump	358
3	#pencecard	1,341	18	#freeassange	356
4	#maga	881	19	#punkaf	354
5	#fightback	702	20	#godwins	244
6	#1776again	672	21	#execute	241
7	#antifaarefascists	607	22	#covidiot	231
8	#blmareracists	607	23	#arrest	228
9	#covid19	606	24	#trampicantraitors	225
10	#treason	555	25	#brandonbernard	223
11	#vote	498	26	#mcenemy	218
12	#trump	452	27	#moscowmitch	215
13	#trump2020	434	28	#againsttrump	199
14	#walterreed	428	29	#makeassholegoaway	199
15	#savebrandonbernard	421	30	#jesuschrist	187

Then, how frequent are violent tweets over time? Figure 3 illustrates the timeline of tweets containing violent political rhetoric. The trend is expressed in terms of their count and of their proportion to the total number of political-keyword tweets. Regardless of the metric, the figure shows very similar trends. First of all, we can see that the proportion of violent political rhetoric is quite rare. For the period of data collection, an average of 0.07% of the tweets that include political keyword(s) contain violent political rhetoric. Such rarity is consistent with findings from recent research on aggressive political communication on social media. For instance, Siegel et al. (2021) reports that around 0.2% of political tweets

contain hate speech during the period from June 17 2015 to June 15 2017. Although violent tweets comprise only a small fraction of political discussion, it is important to note that it amounts to hundreds of thousands of tweets containing violent political rhetoric, per day, and it is seen by the number of users engaged in political communication that is far greater than the number of such tweets themselves.[14]

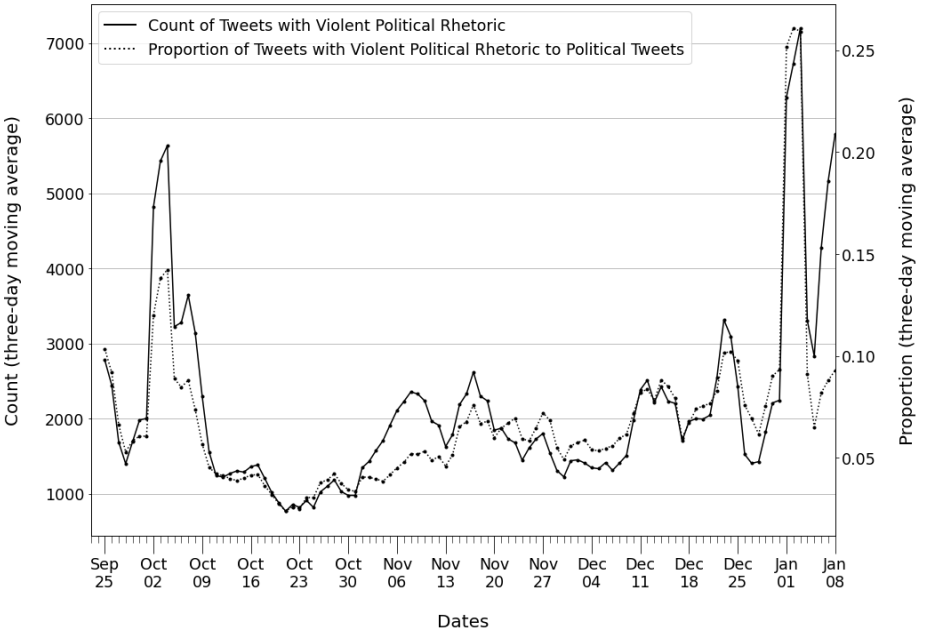


Figure 3. Timeline of violent political rhetoric (September 23 2020 - January 8 2021)

Note: The *y*-axis on the left side indicates the number of tweets containing violent political rhetoric while the other *y*-axis on the right-side depicts the proportion of such tweets relative to tweets containing a political keyword. Each point in the lines indicates the three-day moving average.

As illustrated in Figure 3, there is a considerable over-time variation in the trend of

[14]Note that the Streaming API returns 1% of all tweets in real-time. Therefore, the estimated number of violent tweets will be more than roughly 100 times greater than what we see in the data.

violent political rhetoric. In particular, two big spikes are prominent in early October 2020 and early January 2021 along with a steady increase toward the election and the period of power transition. To provide a detailed look into issues driving the trend, Table 2 reports the weekly top-5 hashtags included in violent tweets. While the steady uptrend toward the election and the period of power transition appears associated with the partisan competition/tension over the election and its results (#vote, #trump2020, #electionday, #laptopfromhell, #tonybobulinski), the two big spikes require further explanation. First, the hashtags for the week from September 30 to October 6 (e.g., #walterreed, #trump, #covidiot, #covid19) show that the earlier spike reflects political animosity surrounding Trump's infection of COVID-19 and his much-criticized behavior during his three-day hospitalization at Walter Reed military (O'Donnell 2020). In addition, manual reading of the tweets on October 2 and the following several days verifies that there were numerous tweets expressing a violent intention against Trump.

As for the later spike, the hashtags for the last couple of weeks, such as #fightback, #1776again, and #pencecard, are the ones that grew substantially among far-right extremists and conspiracy theorists who attempt to delegitimize the election results. We can also see that anti-Trump users, in turn, responded to the far-right discourse using hashtags such as "#arrest and #execute #traitortrump", leading to the massive upsurge in the amount of violent political rhetoric during the last phase of the data collection period (for more detailed information about the context in which these hashtags were used, see Blumenthal 2021, Itkowitz and Dawsey 2020, and Lang et al. 2021). It is also important to note that, while the general prevalence of violent political rhetoric in November and December reflects the partisan tension over the election results (#treason, #diaperdon, #fightbacknow, #stopthesteal) along with other politically salient issues, the drastic uptrend starting in the last week of 2020 appears to be predominantly driven by the extremist discourse agitated

by Trump's continuous mobilizing effort, inside and outside Twitter. Considering Trump's tweet instigating his radical supporters to gather in D.C. on January 6 and the riot on that day,[15] it is abundantly clear that offline political conflict is intertwined with violent political rhetoric on Twitter.[16]

[15]On December 26 2020, Trump tweeted that *"The "Justice" Department and the FBI have done nothing about the 2020 Presidential Election Voter Fraud, the biggest SCAM in our nation's history, despite overwhelming evidence. They should be ashamed. History will remember. Never give up. See everyone in D.C. on January 6th."* On January 6 2021, a joint session of Congress was scheduled to be held to count the Electoral College and to formalize Biden's victory.

[16]It is important to note that I am not making causal claims between violent political rhetoric online and offline political conflict. This is an important direction for future research. For existing works, on the relationship between the two, see Chan, Ghose, and Seamans (2016), Gallacher (2021), Gallacher, Heerdink, and Hewstone (2021), Gallacher and Heerdink (2021), Klein (2019), Mooijman et al. (2018), Olteanu et al. (2018), Siegel (2020), Vegt et al. (2019), and Wei (2019).

TABLE 2 *Most frequent hashtags in violent political rhetoric (weekly)*

	(2020) 9/23-9/29	9/30-10/6	10/7-10/13	10/14-10/20
1	#trump2020	#covid19	#executed	#treason
2	#maga	#vote	#amendments	#biden
3	#treason	#walterreed	#bancapitalisim	#sealteam6
4	#debates2020	#trump	#constitution	#hillaryclinton
5	#whenthesecondwavehits	#covidiot	#government	#obama
	10/21-10/27	10/29-11/3	11/4-11/10	11/11-11/17
1	#crimesagainstchildren	#endnigeria	#jesuschrist	#antifaarefascists
2	#crimesagainsthumanity	#endsars	#trump2020	#blmareracists
3	#laptopfromhell	#vote	#trump	#marchfortrump
4	#tonybobulinski	#trump2020	#maga	#trump2020
5	#moscowmitch	#electionday	#trumpcrimefamily	#treason
	11/18-11/24	11/25-12/1	12/2-12/8	12/9-12/15
1	#treason	#maga	#treason	#savebrandonbernard
2	#maga	#diaperdon	#magabusmusts	#brandonbernard
3	#scif	#fightbacknow	#magaqueentrains	#gopisover
4	#trump	#richardmoore	#bidencheated2020	#abolishthedeathpenalty
5	#democracydemandsit	#headsmustroll	#kag2020	#treason
	12/16-12/22	12/23-12/29	12/30-1/5 (2021)	1/6-1/8
1	#pardonsnowden	#wethepeople	#fightback	#maga
2	#freeassange	#1	#1776again	#traitortrump
3	#punkaf	#pencecard	#godwins	#execute
4	#wethepeople	#pardonsnowden	#divinetiming	#arrest
5	#stopthesteal	#freeassange	#trustgod	#trampicantraitors

Politicians in Violent Tweets

Then, what politicians are mentioned in violent tweets? Tweets can "mention" an account either by directly including it in its text or by replying to tweets written by the account (Twitter 2021a). Table 3 reports what politicians' accounts are mentioned in violent tweets

and presents them by the type of position, political party, and gender. Each cell records the average number of violent tweets that mention politicians' accounts in a given category. First of all, the table shows that Trump is at the center of violent partisan expressions on Twitter. As a single political figure, he appears in far more violent tweets than all the other political accounts combined. Pence, the former vice president, attracts the second largest number of violent tweets followed by the contender for the presidency, Biden, and by the vice-presidential candidate from the Democratic Party, Harris. Also, representatives, compared to governors and senators, receive a small amount of attention in violent political tweets. Presumably, it might be due to the large number of representatives that make them less likely to get sufficient individualized attention to stimulate violent partisan expressions.

Given that Trump (Republican and man) can obscure the comparison based on political party and gender, statistics for political party and gender is reported without violent tweets that mention his account. The second part of the table shows that Republicans appear more frequently than non-Republicans (Democrats and a handful of independent/minor-party politicians). Also, we can see that, on average, men politicians appear more frequently in violent tweets than women politicians.

TABLE 3 *Mean mention count*

		Mention Count
Position	Trump (incumbent president)	137,475
	Pence (incumbent vice president)	18,506
	Biden (candidate for presidency)	8,759
	Harris (candidate for vice presidency	467
	Governors	165
	Senators	479
	Representatives	56
Party	Republican	257
	Non-Republican	117
Gender	Women	103
	Men	207

To further explore how political party, gender, and the type of position correlate with the mentioning of politicians in violent tweets,[17] Table 4 reports the results from a negative binomial regression where the count of mentions in violent tweets, the outcome variable, is regressed against the type of position, political party, and gender. In line with the literature (Southern and Harmer 2019), I include the number of followers to consider the amount of attention given to each politician. To prevent a tiny subset of the observations from being

[17]While mentioning does not necessarily indicate targeting, mentioning is a good proxy for targeting and is often used to measure targeting in the literature (Munger 2021; Siegel et al. 2021). To evaluate the extent to which mentioning indicates targeting in my data, I manually labeled a random sample of 500 tweets taken from the entire data of violent tweets in terms of whether a violent tweet is targeting the mentioned politician. Specifically, I labeled a tweet as relevant when a) the intention of violence in the tweet is targeted at a specific politician, b) the tweet mentions target the politician using their account (mention). The result shows that about 40% of violent tweets target politicians using their accounts. This proportion is higher than what is found in a similar study on hate speech in online political communication (about 25% in Siegel et al. 2021).

overly influential, I exclude the candidates for the presidential election (Biden, Trump, Harris, Pence) who attracted so much attention during the period around the election. For the details of modeling and robustness analysis, see Appendix F.

First, the results reveal that being Republican correlates positively with mentioning in violent tweets (Model 5). Why do Republican politicians appear more frequently in violent tweets than Democratic ones? One possibility is that politicians who belong to the party holding presidency are more frequently targeted as they might draw more attention and criticism, particularly given the amount of violent intention directed at Trump. Also, as often pointed out in the literature, Twitter users are younger and more likely to be Democrats than the general population (Wojcik and Hughs 2019). Therefore, liberal users who outnumber conservative ones might write more violent tweets that target Republican politicians than their conservative counterparts do against Democratic politicians. Second, the results show that being a woman is positively associated with mentioning in violent tweets (Model 5). This is consistent with both academic and journalistic evidence for online abuse against women politicians (Cohen 2021; Di Meco and Brechenmacher 2021; Felmlee, Rodis, and Zhang 2020; Fuchs and SchÄfer 2019; Rheault, Rayment, and Musulan 2019; Southern and Harmer 2019).

TABLE 4 *Mentioning of political accounts: negative binomial regression*

	Model 1	Model 2	Model 3	Model 4	Model 5
Position:governor	1.08***				0.51*
	(0.30)				(0.22)
Position:senator	2.15***				0.18
	(0.23)				(0.18)
Woman		−0.38			0.97***
		(0.21)			(0.15)
Republican			0.78***		0.99***
			(0.18)		(0.13)
Follower Count (log)				2.57***	2.52***
				(0.11)	(0.13)
(Intercept)	4.02***	5.00***	4.47***	−8.79***	−9.44***
	(0.10)	(0.10)	(0.12)	(0.52)	(0.59)
AIC	5255.01	5364.26	5348.76	4689.54	4636.13
BIC	5272.50	5377.38	5361.88	4702.53	4666.46
Log Likelihood	−2623.51	−2679.13	−2671.38	−2341.77	−2311.07
Deviance	734.63	747.13	745.37	664.94	658.53
Number of Accounts	585	585	585	562	562

* Statistical significance: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

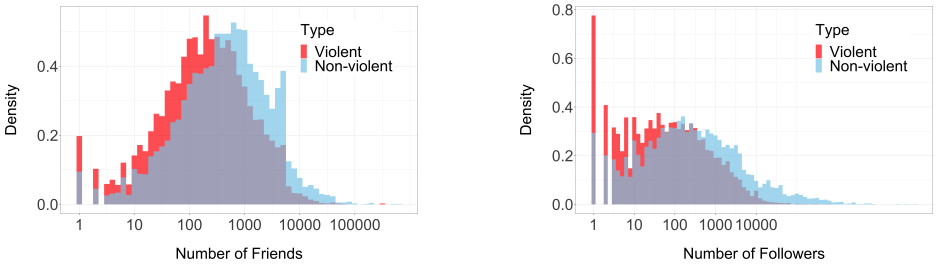
* For Models 4 and 5, the follower count was not retrieved for some accounts due to screen name change, suspension, etc.

Engagement in Political Communication Network by Tweeter Type

How central and active are violent and non-violent users in the political communication network on Twitter? This question is important because the more central to the network

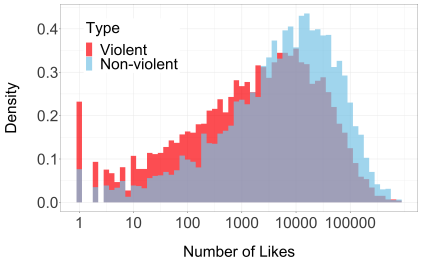
and active violent users are, the more likely ordinary users are exposed to violent political rhetoric. Figure 4 depicts the logged distribution of four user-level indicators in the political communication network (see Appendix G for the median values). Here, violent users follow (and are followed by other users), "like" others' tweets, and write tweets to a lesser degree than non-violent users, implying that violent users are on the fringe in the communication network (the number of friends and followers, and likes) and less active (the number of tweets).[18]

[18]The term "fringe" is used to indicate that violent users are less central in key communication ties on Twitter (i.e., friending, following, and liking). It is most closely related to "degree centrality" in network analysis (i.e., the number of ties that a node has) (Newman 2018).

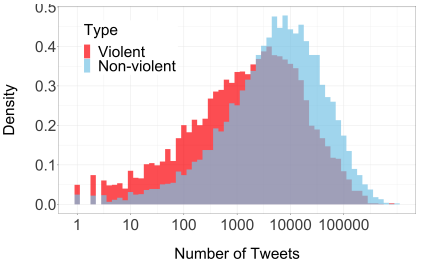


(a) Number of Friends

(b) Number of Followers



(c) Number of Likes



(d) Number of Tweets

Figure 4. Distribution for network engagement indicators

Note: The unit of observation is an account. Each of the four network engagement indicators is depicted on the *x*-axis. The original linear distribution for each indicator was log-transformed (base 10) after adding 1 in order to clearly visualize outliers. The *y*-axis depicts the probability density. "Friends" are whom a given user follows and "followers" are those who follow a given user.

Distribution of Ideology by Tweeter Type

While there is plenty of evidence that far-right extremism is more responsible for offline political violence in the U.S. than their left-wing counterpart (e.g., Jones 2020), it is unclear whether such asymmetry holds in online political communication. How are Twitter users who threaten and incite political violence distributed on the ideological continuum? In Panel (a) in Figure 5, I report the distribution of an ideology score for

violent and non-violent tweeters, measured using an ideal point estimation approach introduced by Barberá (2015).[19] Here, higher scores indicate greater conservativism. First, the distribution of non-violent tweeters shows that they are slightly more liberal (since the vast majority of political tweeters are non-violent ones, the distribution of non-violent tweeters is nearly identical to that of political tweeters). This is consistent with the fact that Twitter users tend to be liberal, younger, and Democrats (Wojcik and Hughs 2019). Second, it is noteworthy is that violent tweeters are more liberal than non-violent tweeters. We can see that the mean ideology score of violent tweeters leans toward the liberal direction. The results of Welch two-sample t-test also show that the difference is 0.18 and statistically significant (95% C.I.: 0.15, 0.20). This analysis reveals that liberals are no less violent than conservatives in online political communication, in contrast to the asymmetry in the offline world.

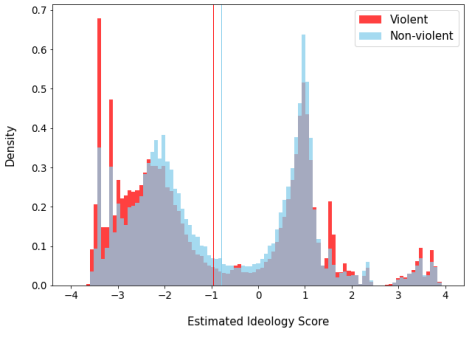
Certainly, the liberal slant might be affected by the fact that the data covers a period that only includes a Republican president. Indeed, a huge number of threatening tweets were targeted at Trump (see Table 3). Considering the level of hostility an incumbent president can provoke from the partisan opposition, liberals might be over-represented in violent tweets in the data. However, the liberal slant still exists after removing all the tweets that mention Trump's account (see Appendix H). Although the difference decreases

[19]This method is well established and has been used in many other studies both in political science and in other social science disciplines (Brady et al. 2017; Freelon and Lokot 2020; Gallego et al. 2019; Hjorth and Adler-Nissen 2019; Imai, Lo, and Olmsted 2016; Jost et al. 2018; Kates et al. 2021; Munger 2021; Sterling, Jost, and Bonneau 2020; Vaccari et al. 2015). It is based on the assumption that Twitter users follow political actors (politicians, think tanks, news outlets, and others) whose positions on the latent ideological dimension are similar to theirs (i.e., homophily in social networks). Considering following decisions a costly signal about users' perceptions of both their latent ideology and that of political actors, the method estimate ideal points of Twitter users based on the structure of the following ties (without any manually labeled data). It produces a uni-dimensional score in which negative values indicate liberal ideology and positive values indicate conservative ideology.

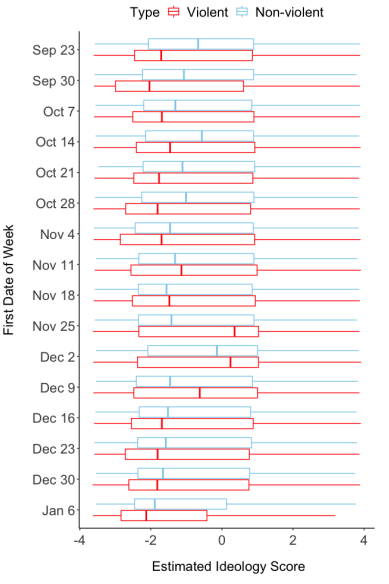
to 0.09 on the continuum, violent users still tend to be more liberal than non-violent users at a statistically significant level (95% C.I.: 0.05, 0.13).

Here, it is important to note that there is over-time heterogeneity. Panel (b) in Figure 5 shows that, while violent users tend to be more liberal than non-violent ones for the first seven weeks, the trend flips for the next five weeks, and again flips back for the last four weeks. These findings imply that the use of violent language in online political communication is likely to reflect how particular phrases of politics stimulate violent partisan hostility — as seen in the hashtags in Table 2 — rather than the use of violent political rhetoric bears an inherent relationship with ideology.

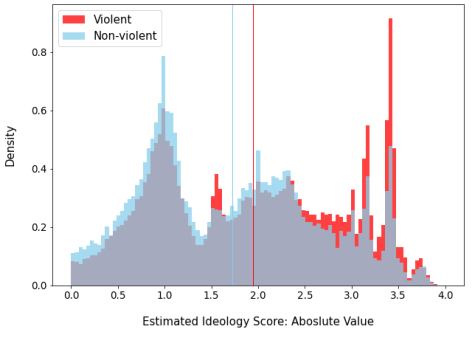
Finally, to get a sense of how ideologically extreme violent users are compared to non-violent users, I computed an ideological extremity score by taking the absolute value of the ideology score. Panel (c) in Figure 5 demonstrates that violent tweeters are more ideologically extreme than non-violent tweeters. The same pattern is also found for almost all of the weekly distributions shown in Panel (d). These results make intuitive sense in that those who display such radical online behavior are unlikely to be ideologically moderate just like offline political violence is committed by extremists on the far ends of the ideological spectrum.



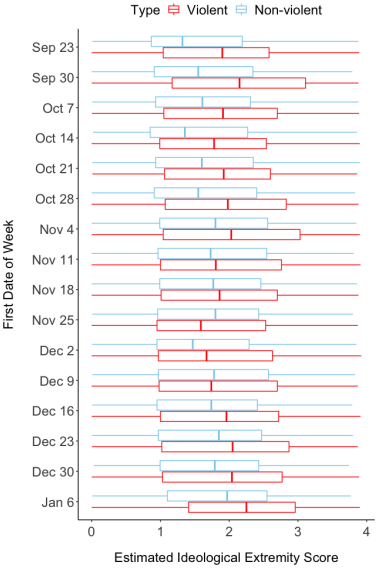
(a) Ideology



(b) Ideology (weekly)



(c) Ideological extremity



(d) Ideological extremity (weekly)

Figure 5. Ideology and ideological extremity by type of political tweeters

Note: The unit of observation is an account. For Panels (a) and (b), larger values indicate greater conservatism. For Panels (c) and (d), larger values indicate greater extremity. The vertical lines in Panels (a) and (c) indicate the mean value for each group.

Spread of Violent Political Rhetoric

How do tweets containing violent political rhetoric spread and how far? Existing research on online political communication suggests that, while political information is exchanged primarily among individuals who are ideologically similar (Barberá et al. 2015), there is also a significant amount of cross-ideological communication (Bakshy, Messing, and Adamic 2015; Barberá 2014). Then, in terms of retweeting, do violent tweets spread primarily among ideologically homogeneous users?[20] The first two panels in Figure 6 present two scatter plots for violent and non-violent tweets where tweeter's ideology score is on the x-axis and retweeters' is on the y-axis. We can clearly see the retweets are highly concentrated in the areas of similar ideology scores. The Pearson's R scores are around 0.7 (0.696 for the violent, 0.713 for the non-violent).

While the findings confirm that retweeting, both violent and non-violent, is affected by ideological homophily, there is a substantial amount of cross-ideological spread in both types of political communication (expressed on the top-left and bottom-right side of the plots). Although the spread of violent political rhetoric takes place primarily among ideologically similar users, the findings imply that users encounter and spread partisan opponents' violent behavior, potentially co-radicalizing each other by feeding off political opponents' violent behavior (Ebner 2017; Knott, Lee, and Copeland 2018; Pratt 2017; Moghaddam 2018).

[20]In the data set, approximately 53% of (non-violent) political tweets are original tweets while the proportion is 75% for violent political tweets. It implies that violent tweets are retweeted less than non-violent tweets, which is consistent with the finding that violent users have fewer followers and their tweets get fewer likes (see Figure 4). However, even though such rhetoric is not written by those who share it, it is important to note that retweets of violent tweets still contain violent political rhetoric and ordinary users are exposed to and influenced by them.

Then, how far do violent tweets travel on the Twitter communication network?[21] As previously discussed, violent tweeters tend to lie on the fringe of the communication network. However, their content still can travel to a large audience through indirect ties. Panel (c) in Figure 6 describes the distribution of the shortest path distance on the following network for all the retweets of violent and non-violent tweets in the data set. Here, the shortest path distance is the minimum number of following ties necessary to connect two users. The distance is estimated as one if the retweeter is in the tweeter's followers list (or the tweeter is in the retweeter's friends list). In a similar manner, the distance is estimated as two if the intersection between the retweeter's friends list and the tweeter's follower list is not an empty set (and if there is no direct follower/following relationship). If neither condition is met, the shortest distance is estimated as three or more.

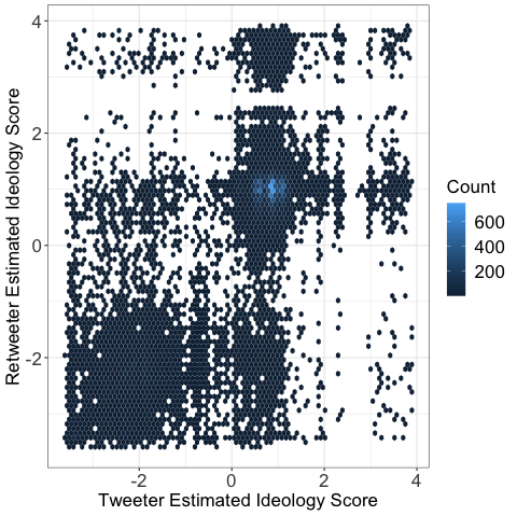
As shown in Panel (c), for both violent and non-violent tweets, around two-thirds of the retweets take place between pairs of users with a direct tie (62% and 67%, respectively). However, there is a substantial minority of retweets that travel beyond the tweeter's followers. Around one thirds of the retweets take place between users whose estimated shortest path distance is two (31% and 27%). For the rest, tweets were retweeted over three or more ties (7% and 6%). The figure shows that political tweets in general spread widely and that violent tweets appear to spread just as far as non-violent tweets despite the offensive nature of the content.[22] Importantly, the findings imply that even if users

[21] It is important to note that not all retweets take place through spread of information on the following network although most retweets take place between connected users on the network (Fábrega and Paredes 2012, 2013). For instance. Twitter has various affordances that enable users to connect with each other. For instance, users can simply search content and other users (Twitter 2021e). Similarly, Twitter's algorithms provide users with popular topics or news, tailored based on who they follow, their interests, and their location (Twitter 2021g).

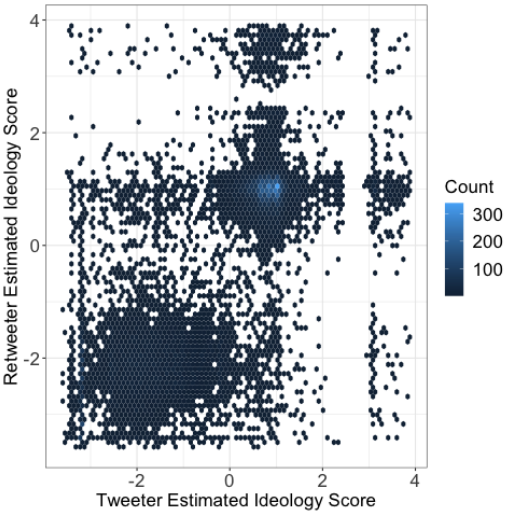
[22] Research on information diffusion in online platforms demonstrates that diffusion between users who are not directly connected is rare and becomes even rarer as the social distance between them increases. Working on seven different online diffusion networks (including the diffusion of

do not follow a violent tweeter (even a violent tweeter's followers), it is still possible that they get exposed to such discomforting content against one's intent. Also, the impact of violent tweets can be dramatically amplified beyond the personal follower networks of violent tweeters if highly popular users, themselves not one of them, retweet a violent tweet, thereby exposing a large number of users to it.

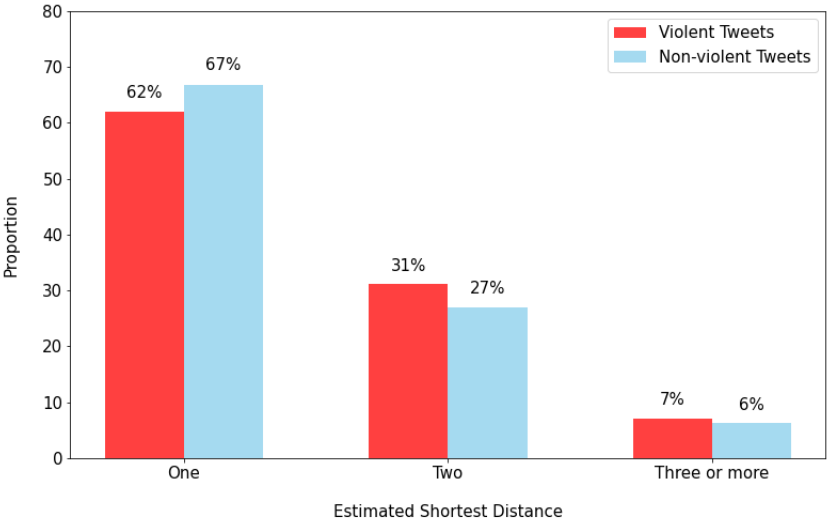
U.S. news articles and YouTube videos on Twitter), Goel, Watts, and Goldstein (2012) finds that approximately 90% of the diffusion takes place between directly connected users. Similarly, focusing on randomly sampled retweets (worldwide), Fábrega and Paredes (2012) reports that over 80% of retweets are between directly connected users and approximately 7% of retweets are between pairs of users whose following distance is two. Retweeting beyond more than two hops on the following network was less than 2%. Given this, relative to general sharing behavior, the results in Panel (c) in Figure 6 provide evidence that the retweeting of violent tweets is generally far-reaching.



(a) Ideological homophily (violent)



(b) Ideological homophily (non-violent)



(c) Reach of political tweets on the following network

Figure 6. Spread of tweets containing violent political rhetoric
Note: For Panels (a) and (b), each point in the plots expresses the number of retweets where the ideology scores of the tweeter and the retweeter correspond to the x-y coordinate. Higher values indicate greater conservatism. For Panel (c), the height of the bars depicts the proportion of tweets containing violent political rhetoric whose shortest distance on the following network belongs to each category. For non-violent tweets, I use a random sample of 238 tweets due to a heavy limit on retrieving follower IDs in the Twitter API (Twitter 2021d).

Conclusion

The recent violent hostility among ordinary American partisans, as dramatically expressed in the Capital Riot, has drawn immense attention both from the media and academia. While the previous literature tends to view partisanship positively as guidance for policy stance and vote choice (Campbell et al. 1980), such view is increasingly replaced by concerns about its destructive potential. At the same time, despite the clear benefits of social media for political outcomes such as political learning and participation (Dimitrova et al. 2014; Tucker et al. 2017), social media platforms are criticized and scrutinized for hateful and violent political communication and their role in stimulating and exacerbating offline violence between confronting partisans.

This paper is among the first to make sense of violent partisan hostility expressed online and thus contribute to the fields of grassroots political violence, online political communication, and violent partisanship. Methodologically, I introduce a new automated method that identifies violent political rhetoric from a massive stream of social media data, adding to the toolkit for measuring violent partisanship. Substantively, I demonstrate that violent political rhetoric on Twitter peaks in the days preceding the Capitol Riot, revealing its close relationship with contentious political events offline. Also, users who threaten violence are ideologically extreme and located on the fringe of the communication network. In terms of targeting, violent tweets are more frequently targeted at women and Republican politicians. While the number of violent tweets is small, such tweets often transcend direct inter-personal connections on the communication network, amplifying their negative effects. Finally, such tweets are shared not only among like-minded users but also across the ideological divide, creating potential for co-radicalization where ideologically extreme users further radicalize each other (Ebner 2017; Knott, Lee, and

Copeland 2018; Moghaddam 2018; Pratt 2017).

In addition, the findings in this paper call for further research on the causes and consequences of violent political rhetoric. First, what are the causal relationships between violent political rhetoric online and offline political violence? While this paper presents abundant evidence for close relationships between the two, it is pressing for future research to scrutinize under what conditions individuals are stimulated to engage in violent acts against out-partisans (both online and offline) and whether/how online and offline violent acts stimulate each other. Second, while recent research in political communication investigates the consequences of exposure to mildly violent political metaphors (Kalmoe 2013, 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019), little attention has been paid to an extreme form of violent language such as threats of violence. Therefore, it is crucial to investigate the effects of exposure to threats of political violence. Does exposure to threatening social media posts have a contagion effect where exposed individuals come to endorse and use violent language? Alternatively, does it stimulate any corrective effort where individuals who encounter such norm-violating behavior detach themselves from negative partisanship?

References

Abramowitz, Alan I, and Kyle L Saunders. 2008. "Is polarization a myth?" *The Journal of Politics* 70 (2): 542–555.

Abramowitz, Alan I, and Steven Webster. 2016. "The rise of negative partisanship and the nationalization of US elections in the 21st century." *Electoral Studies* 41:12–22.

Abramowitz, Alan I, and Steven W Webster. 2018. "Negative partisanship: Why Americans dislike parties but behave like rabid partisans." *Political Psychology* 39:119–135.

Anderson, Craig A, and Brad J Bushman. 2002. "Human aggression." *Annual review of psychology* 53 (1): 27–51.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348 (6239): 1130–1132.

Barberá, Pablo. 2014. "How social media reduces mass political polarization. Evidence from Germany, Spain, and the US." *Job Market Paper, New York University* 46.

———. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23 (1): 76–91.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological science* 26 (10): 1531–1542.

Barrie, Christopher, and Justin Chun-ting Ho. 2021. "academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint." *Journal of Open Source Software* 6 (62): 3272. https://doi.org/10.21105/joss.03272. https://doi.org/10.21105/joss.03272.

Berry, Jeffrey M, and Sarah Sobieraj. 2013. *The outrage industry: Political opinion media and the new incivility.* Oxford University Press.

Blumenthal, Sidney. 2021. "The martyrdom of Mike Pence." *The Guardian* (February). https://www.theguardian.com/commentisfree/2021/feb/07/mike-pence-donald-trump-republicans-religion-evangelical.

Borum, Randy. 2011a. "Radicalization into violent extremism I: A review of social science theories." *Journal of strategic security* 4 (4): 7–36.

———. 2011b. "Radicalization into violent extremism II: A review of conceptual models and empirical research." *Journal of strategic security* 4 (4): 37–62.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences* 114 (28): 7313–7318.

Brice-Saddler, Michael. 2019. "A man wrote on Facebook that AOC 'should be shot,' police say. Now he's in jail." *The Washington Post* (August). https://www.washingtonpost.com/politics/2019/08/09/man-said-aoc-should-be-shot-then-he-said-he-was-proud-it-now-hes-jail-it/.

Broockman, David, Joshua Kalla, and Sean Westwood. 2020. "Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not."

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E Stokes. 1980. *The american voter.* University of Chicago Press.

Chan, Jason, Anindya Ghose, and Robert Seamans. 2016. "The internet and racial hate crime: Offline spillovers from online access." *MIS Quarterly* 40 (2): 381–403.

Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. "Antisocial behavior in online discussion communities." In *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 9. 1.

Claassen, Christopher. 2016. "Group entitlement, anger and participation in intergroup violence." *British Journal of Political Science* 46 (1): 127–148.

Cohen, Marshall. 2021. "Capitol rioter charged with threatening to 'assassinate' Rep. Ocasio-Cortez."

Dadvar, Maral, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. "Improved cyberbullying detection using gender information." In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012).* University of Ghent.

Daugherty, Neil. 2019. "Former MLB player Aubrey Huff says he's teaching his children about guns in case Sanders beats Trump." *The Hill* (November). https://thehill.com/blogs/blog-briefing-room/news/472266-former-mlb-player-aubrey-huff-teaching-his-children-how-to-use.

Davidson, Sam, Qiusi Sun, and Magdalena Wojcieszak. 2020. "Developing a new classifier for automated identification of incivility in social media." In *Proceedings of the fourth workshop on online abuse and harms,* 95–101.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated hate speech detection and the problem of offensive language." In *Eleventh international aaai conference on web and social media.*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Di Meco, Lucina, and Saskia Brechenmacher. 2021. "Tackling Online Abuse and Disinformation Targeting Women in Politics."

Dimitrova, Daniela V, Adam Shehata, Jesper Strömbäck, and Lars W Nord. 2014. "The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data." *Communication research* 41 (1): 95–118.

DiPasquale, Denise, and Edward L Glaeser. 1998. "The Los Angeles riot and the economics of urban unrest." *Journal of Urban Economics* 43 (1): 52–78.

Druckman, James, Samara Klar, Yanna Kkrupnikov, Matthew Levendusky, and John Barry Ryan. 2020. "The political impact of affective polarization: how partisan animus shapes COVID-19 attitudes."

Ebner, Julia. 2017. *The rage: The vicious circle of Islamist and far-right extremism.* Bloomsbury Publishing.

Fábrega, Jorge, and Pablo Paredes. 2012. "Three Degrees of Distance on Twitter." *arXiv preprint arXiv:1207.6839.*

———. 2013. "Social contagion and cascade behaviors on Twitter." *Information* 4 (2): 171–181.

Felmlee, Diane, Paulina Inara Rodis, and Amy Zhang. 2020. "Sexist slurs: reinforcing feminine stereotypes online." *Sex Roles* 83 (1): 16–28.

Fiorina, Morris P, and Samuel J Abrams. 2008. "Political polarization in the American public." *Annu. Rev. Polit. Sci.* 11:563–588.

Freelon, Deen, and Tetyana Lokot. 2020. "Russian disinformation campaigns on Twitter target political communities across the spectrum. Collaboration between opposed political groups might be the most effective way to counter it." *Misinformation Review.*

Fuchs, Tamara, and Fabian SchÄfer. 2019. "Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter." In *Japan Forum,* 1–27. Taylor & Francis.

Fujii, Lee Ann. 2011. *Killing neighbors: Webs of violence in Rwanda.* Cornell University Press.

Gallacher, John, and Marc Heerdink. 2021. "Mutual radicalisation of opposing extremist groups via the Internet."

Gallacher, John D, Marc W Heerdink, and Miles Hewstone. 2021. "Online Engagement Between Opposing Political Protest Groups via Social Media is Linked to Physical Violence of Offline Encounters." *Social Media+ Society* 7 (1): 2056305120984445.

Gallacher, John David. 2021. "Online intergroup conflict: How the dynamics of online communication drive extremism and violence between groups." PhD diss., University of Oxford.

Gallego, Jorge, Juan D Martinez, Kevin Munger, and Mateo Vásquez-Cortés. 2019. "Tweeting for peace: Experimental evidence from the 2016 Colombian Plebiscite." *Electoral Studies* 62:102072.

Gervais, Bryan T. 2015. "Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment." *Journal of Information Technology & Politics* 12 (2): 167–185.

———. 2019. "Rousing the Partisan Combatant: Elite Incivility, Anger, and Antideliberative Attitudes." *Political Psychology* 40 (3): 637–655.

Gill, Paul, John Horgan, and Paige Deckert. 2014. "Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists." *Journal of forensic sciences* 59 (2): 425–435.

Goel, Sharad, Duncan J Watts, and Daniel G Goldstein. 2012. "The structure of online diffusion networks." In *Proceedings of the 13th ACM conference on electronic commerce,* 623–638.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297.

Guess, Andrew, Kevin Munger, Jonathan Nagler, and Joshua Tucker. 2019. "How accurate are survey responses on social media and politics?" *Political Communication* 36 (2): 241–258.

Guynn, Jessica. 2021. "'Burn down DC': Violence that erupted at Capitol was incited by pro-Trump mob on social media." *USA Today* (February). https://www.usatoday.com/story/tech/2021/01/06/trump-riot-twitter-parler-proud-boys-boogaloos-antifa-qanon/6570794002/.

Hammer, Hugo L, Michael A Riegler, Lilja Øvrelid, and Erik Velldal. 2019. "THREAT: A Large Annotated Corpus for Detection of Violent Threats." In *2019 International Conference on Content-Based Multimedia Indexing (CBMI),* 1–5. IEEE.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques.* Elsevier.

Hayes, Philip J, and Steven P Weinstein. 1990. "CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories." In *IAAI,* 90:49–64.

Henson, Billy, Bradford W Reyns, and Bonnie S Fisher. 2013. "Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization." *Journal of Contemporary Criminal Justice* 29 (4): 475–497.

Hjorth, Frederik, and Rebecca Adler-Nissen. 2019. "Ideological asymmetry in the reach of pro-Russian digital disinformation to United States audiences." *Journal of Communication* 69 (2): 168–192.

Horowitz, Donald L. 1985. *Ethnic groups in conflict.-Berkeley, CA: Univ.*

Huber, Gregory A, and Neil Malhotra. 2017. "Political homophily in social relationships: Evidence from online dating behavior." *The Journal of Politics* 79 (1): 269–283.

Humphreys, Macartan, and Jeremy M Weinstein. 2008. "Who fights? The determinants of participation in civil war." *American Journal of Political Science* 52 (2): 436–455.

Hutchens, Myiah J, Jay D Hmielowski, and Michael A Beam. 2019. "Reinforcing spirals of political discussion and affective polarization." *Communication Monographs* 86 (3): 357–376.

Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. "Fast estimation of ideal points with massive data." *American Political Science Review* 110 (4): 631–656.

Isbister, Tim, Magnus Sahlgren, Lisa Kaati, Milan Obaidi, and Nazar Akrami. 2018. "Monitoring targeted hate in online environments." *arXiv preprint arXiv:1803.04757*.

Itkowitz, Colby, and Josh Dawsey. 2020. "Pence under pressure as the final step nears in formalizing Biden's win." *The Washington Post* (December). https://www.washingtonpost.com/politics/pence-biden-congress-electoral/202.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual Review of Political Science* 22:129–146.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, not ideologya social identity perspective on polarization." *Public opinion quarterly* 76 (3): 405–431.

Jigsaw. 2020. https://jigsaw.google.com/.

Jones, Seth G. 2020. "War Comes Home: The Evolution of Domestic Terrorism in the United States."

Jost, John T, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A Tucker. 2018. "How social media facilitates political protest: Information, motivation, and social networks." *Political psychology* 39:85–118.

Kalmoe, Nathan P. 2013. "From fistfights to firefights: Trait aggression and support for state violence." *Political Behavior* 35 (2): 311–330.

———. 2014. "Fueling the fire: Violent metaphors, trait aggression, and support for political violence." *Political Communication* 31 (4): 545–563.

———. 2019. "Mobilizing voters with aggressive metaphors." *Political Science Research and Methods* 7 (3): 411–429.

Kalmoe, Nathan P, Joshua R Gubler, and David A Wood. 2018. "Toward conflict or compromise? how violent metaphors polarize partisan issue attitudes." *Political Communication* 35 (3): 333–352.

Kalmoe, Nathan P, and Lilliana Mason. 2018. "Lethal mass partisanship: Prevalence, correlates, and electoral contingencies." In *American Political Science Association Conference.*

Kates, Sean, Joshua Tucker, Jonathan Nagler, and Richard Bonneau. 2021. "The Times They Are Rarely A-Changin': Circadian Regularities in Social Media Use." *Journal of Quantitative Description: Digital Media* 1.

Kennedy, M Alexis, and Melanie A Taylor. 2010. "Online harassment and victimization of college students." *Justice Policy Journal* 7 (1): 1–21.

King, Gary, Patrick Lam, and Margaret E Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–988.

Klein, Adam. 2019. "From Twitter to Charlottesville: Analyzing the Fighting Words Between the Alt-Right and Antifa." *International Journal of Communication* 13:22.

Knott, Kim, Benjamin Lee, and Simon Copeland. 2018. "Briefings: Reciprocal Radicalisation." *CREST. Online document https://crestresearch. ac. uk/resources/reciprocal-radicalisation.*

Krippendorff, Klaus. 2018. *Content analysis: An introduction to its methodology.* Sage publications.

LaFree, Gary, and Gary Ackerman. 2009. "The empirical study of terrorism: Social and legal research." *Annual Review of Law and Social Science* 5:347–374.

LaFree, Gary, Michael A Jensen, Patrick A James, and Aaron Safer-Lichtenstein. 2018. "Correlates of violent political extremism in the United States." *Criminology* 56 (2): 233–268.

Lang, Marissa, Razzan Nakhlawi, Finn Peter, Frances Moody, Yutao Chen, Daron Taylor, Adriana Usero, Nicki DeMarco, and Julie Vitkovskaya. 2021. "Identifying far-right symbols that appeared at the U.S. Capitol riot." *The Washington Post* (January). https://www.washingtonpost.com/nation/interactive/2021/far-right-symbols-capitol-riot/.

Linder, Fridolin. 2017. "Improved data collection from online sources using query expansion and active learning." *Available at SSRN 3026393.*

Lytvynenko, Jane, and Molly Hensley-Clancy. 2021. "The Rioters Who Took Over The Capitol Have Been Planning Online In The Open For Weeks." *BuzzFeed* (January). https://www.buzzfeednews.com/article/janelytvynenko/trump-rioters-planned-online?scrolla=5eb6d68b7fedc32c19ef33b4.

MacKuen, Michael, Jennifer Wolak, Luke Keele, and George E Marcus. 2010. "Civic engagements: Resolute partisanship or reflective deliberation." *American Journal of Political Science* 54 (2): 440–458.

Magu, Rijul, Kshitij Joshi, and Jiebo Luo. 2017. "Detecting the hate code on social media." In *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 11. 1.

Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. "Spread of hate speech in online social media." In *Proceedings of the 10th ACM Conference on Web Science,* 173–182.

Matsumoto, David, Mark G Frank, and Hyisung C Hwang. 2015. "The role of intergroup emotions in political violence." *Current Directions in Psychological Science* 24 (5): 369–373.

McGilloway, Angela, Priyo Ghosh, and Kamaldeep Bhui. 2015. "A systematic review of pathways to and processes associated with radicalization and extremism amongst Muslims in Western societies." *International review of psychiatry* 27 (1): 39–50.

Miller, Blake, Fridolin Linder, and Walter R Mebane. 2020. "Active learning approaches for labeling text: review and assessment of the performance of active learning approaches." *Political Analysis* 28 (4): 532–551.

Moghaddam, Fathali M. 2018. *Mutual radicalization: How groups and nations drive each other to extremes.* American Psychological Association.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Mooijman, Marlon, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. "Moralization in social networks and the emergence of violence during protests." *Nature human behaviour* 2 (6): 389–396.

Munger, Kevin. 2017. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39 (3): 629–649.

———. 2021. "Don't@ me: Experimentally reducing partisan incivility on Twitter." *Journal of Experimental Political Science* 8 (2): 102–116.

Newman, Mark. 2018. *Networks.* Oxford university press.

Nikolov, Alex, and Victor Radivchev. 2019. "Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles." In *Proceedings of the 13th International Workshop on Semantic Evaluation,* 691–695.

O'Donnell, Carl. 2020. "Timeline: History of Trump's COVID-19 illness." *Reuters* (October). https://www.reuters.com/article/us-health-coronavirus-trump.

Olteanu, Alexandra, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. "The effect of extremist violence on hateful speech online." In *Twelfth International AAAI Conference on Web and Social Media.*

Pauwels, Lieven JR, and Ben Heylen. 2017. "Perceived group threat, perceived injustice, and self-reported right-wing violence: An integrative approach to the explanation right-wing violence." *Journal of interpersonal violence,* 0886260517713711.

Pilkington, Ed, and Sam Levine. 2020. "'It's surreal': the US officials facing violent threats as Trump claims voter fraud." *The Guardian* (December). https://www.theguardian.com/us-news/2020/dec/09/trump-voter-fraud-threats-violence-militia.

Popan, Jason R, Lauren Coursey, Jesse Acosta, and Jared Kenworthy. 2019. "Testing the effects of incivility during internet political discussion on perceptions of rational argument and evaluations of a political outgroup." *Computers in Human Behavior* 96:123–132.

Pratt, Douglas. 2017. "Islamophobia as Reactive Co-Radicalization." In *Religious Citizenships and Islamophobia,* 85–98. Routledge.

Rheault, Ludovic, Erica Rayment, and Andreea Musulan. 2019. "Politicians in the line of fire: Incivility and the treatment of women on social media." *Research & Politics* 6 (1): 2053168018816228.

Romm, Tony. 2021. "Facebook, Twitter could face punishing regulation for their role in U.S. Capitol riot, Democrats say." *The Washington Post* (January). https://www.washingtonpost.com/technology/2021/01/08/facebook-twitter-congress-trump-riot/.

Scacco, Alexandra. 2010. *Who riots? Explaining individual participation in ethnic violence.* Citeseer.

Schils, Nele, and Lieven JR Pauwels. 2016. "Political violence and the mediating role of violent extremist propensities." *Journal of Strategic Security* 9 (2): 70–91.

Settles, Burr. 2009. “Active learning literature survey.”

Shandwick, Weber. 2019. “CIVILITY IN AMERICA 2019: SOLUTIONS FOR TOMORROW.”

Siegel, Alexandra A. 2020. “Online hate speech.” *Social Media and Democracy: The State of the Field, Prospects for Reform,* 56–88.

Siegel, Alexandra A, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, Joshua A Tucker, et al. 2021. “Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath.” *Quarterly Journal of Political Science* 16 (1): 71–104.

Southern, Rosalynd, and Emily Harmer. 2019. “Twitter, incivility and “everyday” gendered othering: an analysis of tweets sent to UK members of Parliament.” *Social Science Computer Review,* 0894439319865519.

Stenberg, Camilla Emina. 2017. “Threat detection in online discussion using convolutional neural networks.” Master’s thesis.

Sterling, Joanna, John T Jost, and Richard Bonneau. 2020. “Political psycholinguistics: A comprehensive analysis of the language habits of liberal and conservative social media users.” *Journal of personality and social psychology* 118 (4): 805.

Suhay, Elizabeth, Emily Bello-Pardo, and Brianna Maurer. 2018. “The polarizing effects of online partisan criticism: Evidence from two experiments.” *The International Journal of Press/Politics* 23 (1): 95–115.

Sydnor, Emily. 2019. *Disrespectful democracy: The psychology of political incivility.* Columbia University Press.

Tausch, Nicole, Julia C Becker, Russell Spears, Oliver Christ, Rim Saab, Purnima Singh, and Roomana N Siddiqui. 2011. "Explaining radical group behavior: Developing emotion and efficacy routes to normative and nonnormative collective action." *Journal of personality and social psychology* 101 (1): 129.

Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. "The dynamics of political incivility on Twitter." *Sage Open* 10 (2): 2158244020919447.

Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. 2016. "A bad workman blames his tweets: the consequences of citizens' uncivil Twitter use when interacting with party candidates." *Journal of communication* 66 (6): 1007–1031.

Tucker, Joshua A, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. "From liberation to turmoil: Social media and democracy." *Journal of democracy* 28 (4): 46–59.

Twitter. 2021a. "About replies and mentions." Accessed: 25 October 2021. https://help.twitter.com/en/using-twitter/mentions-and-replies.

———. 2021b. "Academic Research product track." Accessed: 25 October 2021. https://developer.twitter.com/en/products/twitter-api/academic-research.

———. 2021c. "Filter realtime Tweets." Accessed: 14 October 2021. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview.

———. 2021d. "GET followers/ids." Accessed: 19 November 2021. https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/api-reference/get-followers-ids.

Twitter. 2021e. “How to use Twitter search.” Accessed: 3 November 2021. https://help.twitter.com/en/using-twitter/twitter-search.

———. 2021f. “Search Tweets: standard v1.1.” Accessed: 14 October 2021. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets.

———. 2021g. “Twitter Trends FAQ.” Accessed: 3 November 2021. https://help.twitter.com/en/using-twitter/twitter-trending-faqs.

———. 2021h. “Violent threats policy.” Accessed: 14 October 2021. https://help.twitter.com/en/rules-and-policies/violent-threats-glorification.

Vaccari, Cristian, Augusto Valeriani, Pablo Barberá, Rich Bonneau, John T Jost, Jonathan Nagler, and Joshua A Tucker. 2015. “Political expression and action on social media: Exploring the relationship between lower-and higher-threshold political activities among Twitter users in Italy.” *Journal of Computer-Mediated Communication* 20 (2): 221–239.

Vegt, Isabelle van der, Maximilian Mozes, Paul Gill, and Bennett Kleinberg. 2019. “Online influence, offline violence: Linguistic responses to the’Unite the Right’rally.” *arXiv preprint arXiv:1908.11599*.

Vigdor, Neil. 2019. “Police officer suggests AOC should be shot: ’She needs a round’.” *Independent* (July). https://www.independent.co.uk/news/world/americas/us-politics/aoc-trump-twitter-democrats-louisiana-police-charlie-rispoli-a9015301.html?utm_source=share&utm_medium=ios_app.

Wang, Shuai, Zhiyuan Chen, Bing Liu, and Sherry Emery. 2016. "Identifying search keywords for finding relevant social media posts." In *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 30. 1.

Waseem, Zeerak, and Dirk Hovy. 2016. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In *Proceedings of the NAACL student research workshop,* 88–93.

Weerkamp, Wouter, Krisztian Balog, and Maarten de Rijke. 2012. "Exploiting external collections for query expansion." *ACM Transactions on the Web (TWEB)* 6 (4): 1–29.

Wei, Kai. 2019. "Collective Action and Social Change: How Do Protests Influence Social Media Conversations about Immigrants?" PhD diss., University of Pittsburgh.

Wester, Aksel, Lilja Øvrelid, Erik Velldal, and Hugo Lewi Hammer. 2016. "Threat detection in online discussions." In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,* 66–71.

Wester, Aksel Ladegård. 2016. "Detecting threats of violence in online discussions." Master's thesis.

Westwood, Sean, Justin Grimmer, Matthew Tyler, and Clayton Nall. 2021. "American Support for Political Violence is Low."

Williams, Matthew L, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. "Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime." *The British Journal of Criminology* 60 (1): 93–117.

Wojcik, Stefan, and Adam Hughs. 2019. “Sizing Up Twitter Users.” Accessed February 21, 2021. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. “Ex machina: Personal attacks seen at scale.” In *Proceedings of the 26th international conference on world wide web,* 1391–1399.

Zeitzoff, Thomas. 2020. “The Nasty Style: Why Politicians Use Violent Rhetoric.” *Unpublished working paper.*

Zimmerman, Steven, Udo Kruschwitz, and Chris Fox. 2018. “Improving hate speech detection with deep learning ensembles.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Appendix

A. Selection of Political Keywords

In this paper, I intend to build a set of political tweets that engage major politicians on Twitter (not a representative sample of "all political tweets."). And, this is why I focus on the "mention" function in Twitter (Twitter 2021a). The mention function is a key feature that structures communication on Twitter by allowing users to connect to each other and stay updated on their conversations with other users. As described in the main text, I use a very broad sample of U.S. politicians' Twitter accounts which include accounts for Republican and Democratic parties as well as ones for members of the Congress, governors, the president, the vice president, and their contenders in the 2020 Presidential Election.

While I "mention" is a central feature around communication on Twitter, one other major way to engage politicians is simply using names (e.g., "Donald Trump" as opposed to "@realDonaldTrump"). While I considered using both accounts and names of the politicians in filtering live tweets, it was impossible due to the limit to the number of keywords that can be included in using the Streaming API (Twitter 2021c). In addition, using names as opposed to accounts is prone to measurement error for many reasons. First, politicians are called by different versions of their names and it is extremely difficult, if not impossible, to decide on a particular form for each and every politician. For instance, there are cases where politicians are called by the last name only, the full name, or various abbreviations (e.g., TJ Cox). Also, there are issues related to homonyms for many politicians and it is practically impossible to separate out tweets referencing other people whose name is the same as politicians of interest (e.g., the North Carolina Representative

David Price and the baseball pitcher David Price).

To determine the full name of politicians, I use the name that appears on a given politician's Wikipedia page. When the Wikipedia page shows a full name including a middle name or an abbreviation, I referred to the politicians' Twitter page and followed the name used there. To count the number of the two groups of tweets, I used an R package `academictwitteR` (Barrie and Ho 2021) and accessed the newly introduced Academic Research API which allows for access to a full archive of tweets beyond the standard seven-day limit (Twitter 2021b). I counted the number of tweets including full name tweets and mention tweets for each day in the data collection period (September 23, 2020 through January 8, 2021) and aggregated them by politicians' accounts.

Figure A1 depicts the distribution of the proportion of the number of full name tweets to the number of mention tweets (expressed in percentage). The original distribution is highly skewed so I log-transformed it. The figure shows that most of the observations are concentrated in the area left to the 100% point at which the numbers of full name tweets and of mention tweets are equal. Because the distribution is skewed, I used the median for the central tendency measure. The median proportion, 13.04%, indicates that only a small fraction of tweets engaging politicians on Twitter use their full names as opposed to their accounts. In addition, Table A1 breaks down the median proportion across the gender, political party, and position of politicians. We can see that there are no noticeable discrepancies in the proportion across the three characteristics. This provides evidence that tweets including politicians' full names and tweets including their accounts are not systematically different with regard to the key dimensions of comparison in the substantive analysis.

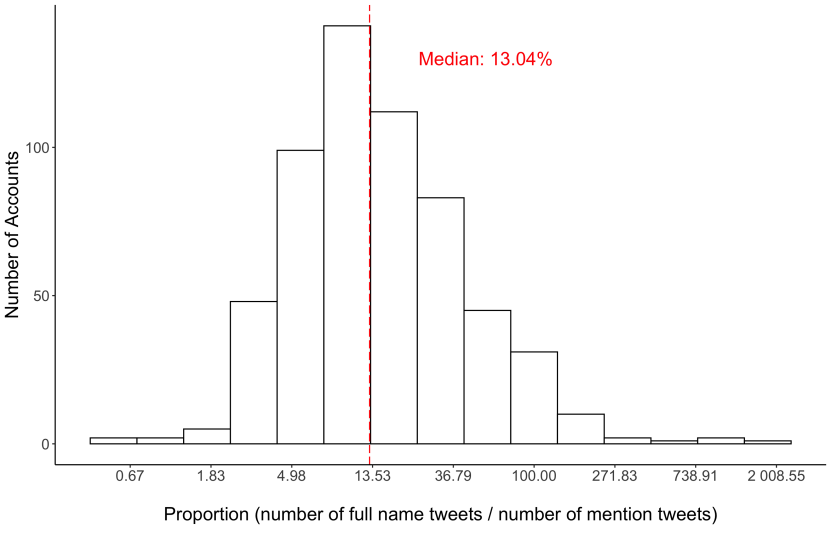


Figure A1. Distribution of the proportion of full-name tweets to mention tweets)

Note: The unit of observation is a politician's account. The x-axis depicts the proportion in percentage. The y-axis is for the count of observations.

TABLE A1 *Median proportion of full-name tweets to mention tweets*

		Proportion
Gender	Women	12.2%
	Men	13.1%
Political Party	Republican	12.8%
	Non-Republican	13.2%
Position	Governors	11.9%
	Senators	13.2%
	Representatives	13.1%
Total		13.0%

B. Selection of Violent Keywords

Keyword filtering (similarly, dictionary methods) is a widely used tool to automate content analysis for large textual data (Grimmer and Stewart 2013). One of the key challenges in using keywords to retrieve relevant documents is to compile a good list of keywords as humans are generally limited in recalling a comprehensive and unbiased list of keywords related to a specific concept (Hayes and Weinstein 1990; King, Lam, and Roberts 2017). Therefore, recent works have focused on developing innovative methods to discover/expand a set of keywords (King, Lam, and Roberts 2017; Wang et al. 2016). My method is in line with a group of methods where researchers rely on an external corpus to expand keywords for document retrieval (Weerkamp, Balog, and Rijke 2012). While none of the keyword expansion methods is capable of retrieving "all" keywords, I effectively draw insights into violent lexical features using an external data set of massive size (approximately two million Wikipedia comments) where human coders label documents for whether a given comment is threatening or not.

It is important to note that the set of keywords I extract from the corpus is highly comprehensive (link to the list of violent keywords and relevant replication materials). I extract 200 uni- and bi-gram keywords that are the most predictive of perceived threat. I experimented with different threshold values (e.g., 100, 200, 300, 400) and set the threshold based on my judgment of the point beyond which keywords are no longer meaningfully associated with perceived threat and are likely to only result in false positives. Since I manually label the resulting violent-keyword tweets (tweets filtered in through the 200 violent keywords) in terms of whether the tweet is actually violent in the next step (Step 3), I was able to include even keywords that are marginally associated with threat perception. The selected keywords not only involve a wide variety of keywords (e.g., die, punch,

choke) but also cover their semantic variants (e.g., “die”, “dead”, “death”). Furthermore, the list involves many keywords that are not necessarily violent themselves but are often used in violent rhetoric (swear words, auxiliary verbs, or collocative structures).

While the list is very broad and there is little reason to believe that missing violent rhetoric would introduce any bias in a predictable manner, it is still important to note that not all tweets containing violent political rhetoric are filtered in with the list of keywords. This is because violent political rhetoric can be used without having any violent keyword (or keywords used in combination with violent keywords), making any keyword approach ineffective. For instance, in texts like “I have my eye on you, so you better watch your back tonight”, each of the words is not particularly violent in meaning but the text still conveys a violent intention. While the list of keywords does include ones that are not violent at all in isolation but still carry a violent intention in context, we cannot be perfectly sure such keywords will perfectly capture tweets where a violent intention is expressed in a subtle way. To the best of my knowledge, this is an area that has not yet been extensively studied in the field of natural language processing and thus requires further work.

C. Manual Labeling

Three human coders (including myself and two undergraduate assistants) labeled tweets in terms of whether a given tweet contains violent political rhetoric or not. Specifically, the human coders were instructed to classify tweets into three classes. Class 1 is “violent politic rhetoric" where the author expresses the intention of physical harm against a political opponent. Class 2 is “violent political metaphor" where the author’s statement about essentially non-violent politics is expressed using a violent metaphor but still lacks a violent intention. While Class 2 is not directly related to my study, tweets that fall into

this class appear frequently enough to constitute a separate class. Class 3 is a garbage can class for tweets that are neither Class 1 nor Class 2. The tweets were presented as reformatted on Google Sheet. A tweet that quotes another tweet is presented with the quoted tweet because the former's meaning is more clear with the latter.

The concept of violence is inherently ambiguous and subjective. Therefore, it was necessary to refine coding rules throughout the manual annotation process. The major sources of false positives involve a) when violent phrases are used as a metaphor that describes non-violent political events as violent (Kalmoe 2013, 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019), b) a religious curse that does not involve any real violence (e.g., 'burn in hell!'), c) quoting (or even criticizing) violent political rhetoric from someone else, and d) irony (e.g., 'why don't just shoot them all if you believe violence solves the problem?'). Click this link for the most detailed coding rules.

The coders manually labeled a set of 2,500 tweets together (meaning each tweet is labeled three times). Specifically, the coders worked together on the initial 2,000 tweets to refine coding guidelines and manually labeled another 500 tweets. Again, after manually annotating the 500 tweets, the coding guidelines were updated. Then, the coding guidelines based on the 2,500 tweets were used for later manual annotation of another set of 7,500 tweets. For the 7,597 tweets, three coders worked on three different sets of tweets (Coder 1: 3,500, Coder 2: 3,500, Coder 3: 597). In sum, a total of 10,097 tweets were manually labeled.

As previously noted, the concept of violent political rhetoric is inherently subjective. Accordingly, the levels of inter-coder agreement for similar studies on aggressive online behavior are low to moderate (Table A2). The inter-coder agreement scores achieved in our study are reported in Table A3. Our study achieves a Krippendorff's Alpha score close to 0.6. It shows that, by any measure, the level of inter-coder agreement outperforms the

standard in the relevant literature.

TABLE A2 *Inter-coder agreement on similar concepts*

Study	Concept	Krippendorff's Alpha
Theocharis et al. (2016)	political incivility	0.54
Munger (2021)	partisan incivility	0.37
Wulczyn, Thain, and Dixon (2017)	personal attacks	0.45
Cheng, Danescu-Niculescu-Mizil, and Leskovec (2015)	antisocial language	0.39

TABLE A3 *Inter-coder agreement on 500 manually-labeled tweets*

Measure	Coder 1&2	Coder 2&3	Coder 1&3
Cohen's Kappa	0.569	0.622	0.593
Light's Kappa		0.595	
Fless's Kappa		0.597	
Krippendorff's Alpha		0.597	

D. Active Learning and Machine Classification

Relying on active learning (Linder 2017; Miller, Linder, and Mebane 2020; Settles 2009), I followed the next process to build training data for my final machine learning classifier.

1. I take a random sample of M tweets from a corpus of tweets containing political and violent keywords (C_{pv}).

2. Including myself, three human annotators label the M tweets in terms of whether a given tweet contains a threat of violence or not. A machine learning classifier is trained on the labeled tweets.

3. Next, the trained classifier is fit on the rest of C_{pv} and the predicted probability of being violent is calculated.

4. I select another (non-random) set of tweets whose probability of belonging to the violent class lies just above or below the decision threshold. These are the tweets whose class the classifier is most uncertain about. The tweets are manually labeled and added to the existing labeled tweets.

5. The process from 2 to 4 is iterated until resources are exhausted and/or the performance of the final classification is satisfying.

For the first round, I randomly sampled 2,500 tweets and labeled them with undergraduate coders. Then, I trained a logistic regression classifier using the count vectors of uni- and bi-grams as features. In the second round, I used the logistic regression classifier to select 7,000 tweets whose probability of belonging to the threat class is around the decision boundary (p = 0.5). Each of the two undergraduate coders labeled 3,500 tweets, independently. In the third round, I fit a fined-tuned BERT (Bidirectional Encoder Representations from Transformers) classifier to select over 500 tweets for additional manual annotation (for detailed information about BERT, see Devlin et al. 2018). Through this iterative process, a total of 10,097 tweets containing political and violent keywords are manually labeled.

With the final training set of N = 10,097, I fit various machine learning classifiers. Since the data set is imbalanced, I used precision, recall, and F-1 score to evaluate the performance of classifiers. To evaluate their performance more reliably, I use K-fold cross validation (K = 5). Here, the training set is randomly partitioned into 5 equally-sized chunks. Out of the 5 chunks, a single chunk is retained as the validation data for testing the model, and the remaining four chunks are used together to build a classifier. This process is repeated five times and the performance is averaged across each validation experiment. The results of the 5-fold cross validation are reported in Table A4 and A5.

As shown in the tables, the BERT model achieves the best performance and is used for final classification. For the BERT model, the binary decision threshold is set at 0.875 since most relevant cases start to appear on the right tail of the probability distribution. The BERT model parallels or outperforms the performance achieved in the relevant literature. When it comes to identifying social media posts involving a threat of violence. A small body of research on YouTube proposes a series of approaches that mainly rely on natural language processing and machine learning, similar to my approach. These works rely on a data set of YouTube comments. The data set, collected by Hammer et al. (2019) in 2013, consists of comments from 19 different YouTube videos concerning highly controversial religious and political issues in Europe. Using the data set, Wester (2016) and Wester et al. (2016) build a series of statistical classifiers with various lexical and linguistic features. They achieve their best performance, using combinations of simple lexical features (F-1: 68.85). Using the same data set, Stenberg (2017) builds various convolutional neural network models and achieves a similar performance (F-1: 65.29).

While the BERT model performs well, it is important to note that the model inevitably makes errors. This is particularly the case when tweets involve many violent expressions related to the use of violence. For instance, discussion of the death penalty (e.g., the case of Brandon Bernard) tends to involve many violent expressions (e.g., kill, death, die, etc) and classifying tweets in this context can be a challenging task for any machine learning model. Obviously, the BERT model still successfully identifies violent political rhetoric arising from such discussion (e.g., *"@realDonaldTrump If you don't stop the execution of Brandon Bernard I hope you die a very painful death. It is the least you deserve you POS!"* or *"@realDonaldTrump @Varneyco I hope you catch an illness and die you orange turd it should've been you and Kyle Rittenhouse that should be injected with poison not Brandon Bernard I hope the White House burns down with you in it"*). At the same time, however, it

can and do misclassify tweets simply discussing (or opposing) the prisoner being executed as violent (e.g., *"@realDonaldTrump BASTARD WHYD U N UR TEAM KILL EXECUTE BRANDON BERNARD"* or *"Brandon Bernard will be executed on HumanRightsDay"*).

TABLE A4 *The average performance of classifiers from 5-fold cross validation*

Model	Precision	Recall	F-1
Logistic Regression + Count Vector	68.51	34.18	45.58
Logistic Regression + TF-IDF Vector	82.06	10.21	18.13
Logistic Regression + GloVe	63.05	11.21	19.01
Random Forest + Count	77.30	19.40	30.94
Random Forest + TF-IDF Vector	81.58	17.38	28.63
Random Forest + GloVe	76.04	10.71	18.77
XGBoost + Count Vector	76.94	7.88	14.24
XGBoost + TF-IDF Vector	77.93	11.54	20.02
XGBoost + GloVe	70.94	14.69	24.31
BERT	71.80	65.61	68.42

TABLE A5 *The results of 5-fold cross validation for the BERT classifier*

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Metric	Precision	72.95	75.98	68.67	72.96	68.42	71.80
	Recall	67.62	57.62	64.78	70.66	67.38	65.61
	F-1	70.18	65.54	66.67	71.79	67.90	68.42
	Accuracy	91.04	90.94	89.80	91.28	89.65	90.54
Count	True Positive	213	174	206	224	221	207.60
	False Positive	79	55	94	83	102.00	82.60
	True Negative	1626	1663	1607	1619	1589	1620.80
	False Negative	102	128	112	93	107	108.40

E. Top-30 Keywords by Type-specificity

Table A6 reports the top-30 keywords that differentiate violent and non-violent political keywords.

TABLE A6 *Comparison of terms by type of tweets*

Rank	Non-violent	Violent	Rank	Non-violent	Violent
1	vote	will	16	senat	@vp
2	georgia	die	17	report	save
3	ballot	@realdonaldtrump	18	legal	arrest
4	presid	hope	19	knew	jail
5	count	penc	20	win	kick
6	elector	execut	21	tax	death
7	great	fuck	22	campaign	go
8	ralli	treason	23	januari	coward
9	voter	like	24	counti	@secpompeo
10	read	@senatemajldr	25	health	trial
11	work	face	26	pennsylvania	hang
12	question	need	27	via	squad
13	ga	fire	28	retweet	await
14	video	ass	29	tweet	sing
15	court	@mike_penc	30	record	@courie85

F. Regression Analysis on Mentioning

Table A6 reports descriptive statistics for the mentioning analysis. Tables A7 and A8 report two additional models to assess whether the findings in the main text are robust to model specifications. The first model is the same as the main model but includes

three candidates for the Presidential Election: Biden, Pence, Harris (except for Trump who is overly influential). The second model is a zero-inflated negative binomial model to account for excess zeros (the first-stage model uses the same set of variables as the second-stage model). Negative binomial regression is used for all of the models to deal with over-dispersion. As seen in the coefficients, the results for position, gender, and partisan affiliation are consistent across the models.

TABLE A7 *Descriptive statistics for mentioning analysis*

Mention Count	Follower Count	Gender	Party	Position
Min. : 0.0	Min. : 2496	Women: 136	D :303	Representative: 436
1st Qu.: 2.0	1st Qu.: 21772	Men: 449	DFL: 1	Governor: 50
Median : 6.0	Median : 37047		I : 2	Senator: 99
Mean : 136.7	Mean : 191013		L : 1	
3rd Qu.: 27.0	3rd Qu.: 105734		R :278	
Max. :25266.0	Max. :12102376			

TABLE A8 *Mentioning/targeting of political accounts: negative binomial regression + Biden/Pence/Harris*

	Coefficient (S.E.)
Position:Biden	−0.30
	(1.44)
Position:Pence	1.19
	(1.42)
Position:Harris	−2.96*
	(1.42)
Position:Governor	0.51*
	(0.22)
Position:senator	0.18
	(0.18)
Women	0.97***
	(0.15)
Republican	0.99***
	(0.13)
Follower Count (log)	2.52***
	(0.13)
(Intercept)	−9.44***
	(0.59)
AIC	4700.63
BIC	4744.00
Log Likelihood	−2340.31
Deviance	662.10
Num. obs.	565

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

TABLE A9 *Mentioning/targeting of political accounts: zero-inflated negative binomial regression*

	Coefficient (S.E.)
Count model: (Intercept)	−9.44***
	(0.56)
Count model: Position:governor	0.49*
	(0.21)
Count model: Position:senator	0.17
	(0.19)
Count model: Women	1.06***
	(0.17)
Count model: Republican	0.99***
	(0.14)
Count model: Follower Count (log)	2.52***
	(0.12)
Count model: Log(theta)	−0.61***
	(0.06)
Zero model: (Intercept)	−0.72
	(97.42)
Zero model: Position:governor	−16.07
	(3332.84)
Zero model: Position:senator	−7.90
	(47.27)
Zero model: Women	11.36
	(97.13)
Zero model: Republican	−1.15
	(2.39)
Zero model: Follower Count (log)	−2.71
	(1.55)
AIC	4639.70
Log Likelihood	−2306.85
Num. obs.	562

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

G. Network Engagement Indicators

Table A10 reports the median value for the four network engagement indicators.

TABLE A10 *Median value for network engagement indicators*

Count	Violent	Non-violent
Friends	199	460
Followers	52	203
Likes	2273	6966
Tweets	1882	5843

H. Distribution of Ideology by Type of Political Tweeters (without Trump's account)

Figure A2 depicts the distribution of ideology by type of tweeters (violent vs. non-violent, without tweets that mention Trump's account. While the gap between the mean ideological scores for the two groups decreases, violent users are still more liberal than non-violent ones.

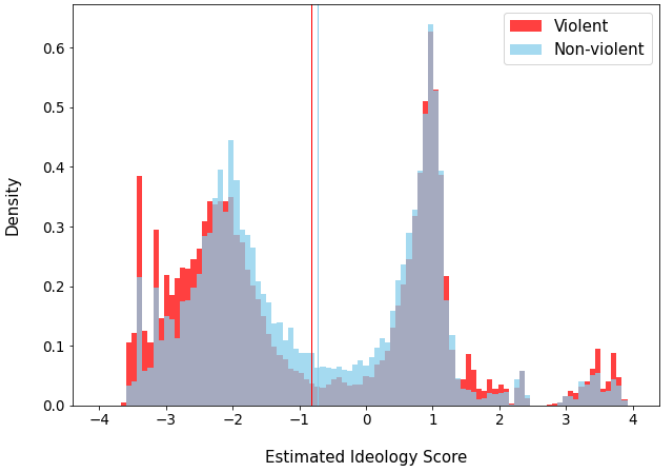


Figure A2. Distribution of ideology by type of political tweeters (without Tweets mentioning '@realDonaldTrump')

Note: The unit of observation is an account. The x-axis depicts the ideology score with larger values indicating greater conservatism. The y-axis is probability density. The vertical lines indicate the mean value for each group.