The European Political Science Association, 2021 *doi:10.1017/psrm.2021.XX*

Violent Political Rhetoric on Twitter [*]

TAEGYOON KIM

Violent hostility between ordinary partisans is undermining American democracy. Social media, once touted as democratizing technology, is blamed for political rhetoric threatening violence against political opponents and implicated in extremist offline political violence. Focusing on Twitter, I propose an automated method to detect such rhetoric and investigate its characteristics. Using a data set spanning a 16-week period surrounding the 2020 Presidential Election, I demonstrate a) violent tweets peak in the days preceding the Capitol Riot, testifying their close relevance to contentious offline politics, b) violent tweets disproportionately target women and Republican politicians, c) violent tweets are rare but spread widely, reaching an audience without a direct tie to violent users, d) violent users are ideologically extreme and on the fringe of the communication network, e) violent tweets are shared primarily among like-minded users but exhibit cross-ideological spread as well.

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

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

early hopes are being overshadowed by mounting concerns about aggressive political communication. In recent days, one can easily encounter uncivil political discussion 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 2018). 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 violence 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 behavioral manifestation of radical/lethal partisanship where individuals not just hate out-partisans (Abramowitz and S. W. Webster 2018) but also support and even enjoy the use of offline 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 respond to offline-world politics? What types of political figures are threatened? What users threaten violence 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 politically 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, focusing 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 analysis on the characteristics and the 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 political elites in the opposition in response to, or exacerbating, contentious political events/issues in the offline world. The online-offline relationship identified in my study opens up a future research agenda on what causal mechanisms connect threats of political violence in online space 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, Gervais 2019, Munger 2017b, Munger 2017c, Popan et al. 2019, Siegel, Tucker, et al. 2019, Siegel 2018, Siegel 2018, Suhay, Bello-Pardo, and Maurer 2018). Given the extremely volatile climate in contemporary American democracy, there is a pressing need to investigate such threatening political language. However, there was little comprehensive empirical analysis on this topic. 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 35% 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 radical/lethal partisanship is manifested online in the form of threats against partisan opponents. Recent studies on mass partisan polarization have highlighted 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 radical/lethal partisanship is expressed in online space. My work contributes to the literature by providing an easy-to-access indicator for tracking the level of radical/lethal partisanship. Considering the evidence that there are significant discrepancies between objective online behavior and survey self-reports (Guess et al. 2019), the current approach is an excellent complement to survey-based measurement as it enables researchers to directly observe the over-time trend of violent partisan behavior expressed in online space. For instance, I illustrate that the level of violent political rhetoric on Twitter corresponds to the violent partisan relations in the offline world, reaching its peak in the days preceding the Capitol Riot.

Related Work

Here, I introduce three strands of literature, each of which provides a context for as well as theoretical insight into the current study on violent political rhetoric on Twitter. First, not only does my study expand the literature on individuals' participation in offline political violence but it also benefits from it. Threatening political violence in online space is itself a violent act and thus can share causal factors with offline political violence. Also, a large body of works taking a micro-level approach to individuals' participation in offline political violence can provide rich theoretical insight into the conditions under which individuals threaten political opponents in online space. Second, although my work is among the first to investigate violent political rhetoric in online space, an extensive body of research investigates politicians' use of violent political metaphors in offline world and aggressive speech in online political discussion, together providing a rich context for the inquiry into violent political rhetoric in online space. Third, violent political rhetoric on social media is a form of behavioral manifestation of extreme negative partisanship. Thus, the literature on partisan polarization and negative partisanship is very useful in understanding why Twitter users express violent intention against out-partisans and what consequences such behavior has.

Offline Political Violence

Although few studies exist to explain political violence in online space, there is an extensive body of literature devoted to explaining why individuals engage in offline political violence in various settings. First, mainly focused on conflict-ridden contexts, a group of works seeks to explain why individuals participate in inter-group violence (ethnic, religious, partisan). Major explanations include selective incentives provided by group leaders that enable individuals to overcome the problem of free-riding (DiPasquale and Glaeser 1998, Humphreys and Weinstein 2008), social pressure from in-group members (Fujii 2011, Scacco 2010), and perceived inequality in the distribution of resources among groups and ensuing anger (Claassen 2016).

Also, an interdisciplinary stream of studies on violent extremism (including criminology and psychology) seek to identify a host of risk factors that are associated with individuals' tendency to join violent extremist activities (Borum 2011a, Borum 2011b, Gill, Horgan, and Deckert 2014, LaFree and Ackerman 2009, McGilloway, Ghosh, and Bhui 2015). Lack of stable employment, history of mental illness, criminal record, low self-control, perceived injustice, and exposure to violent extremism (content and peers) are few among the factors highlighted in the literature (LaFree et al. 2018, Pauwels and Heylen 2017, Schils and Pauwels 2016).

Finally, a recent wave of works in political communication highlights the role of politicians' rhetoric. Kalmoe (2014) finds that exposure to violent political metaphors (metaphors that describe politics as violent events such as a battle or a war) during an electoral campaign increases support for political violence among people with aggressive personalities. Similarly, Matsumoto, Frank, and Hwang (2015) finds that political leaders' rhetoric arousing "ANCODI (anger, contempt, and disgust)" emotions can generate intergroup violence. Focusing on the 2015 Baltimore protests, Mooijman et al. (2018) shows that moralization of political issues can lead to the endorsement of violent protests.

Aggressive Political Communication

Raising concerns about political elites' violent rhetoric in the U.S., a recent stream of studies investigate its consequences for various outcomes. Kalmoe (2019) shows that violent political metaphors 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 the context of political elites' offline speech, many works in online political communication have focused 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 different race, gender, and partisan affiliation (Berry and Sobieraj 2013, Kennedy and Taylor 2010, Munger 2017b, Munger 2017c, Shandwick 2019). Also, they report evidence that such content is becoming more and more prevalent on social media. Aggressive online speech is reported to discourage participation in online discussion (Henson, Reyns, and Fisher 2013), exacerbate inter-group evaluations, and discourage democratic deliberation (Gervais 2019). Accordingly, a large body of works are also devoted to detecting (Davidson et al. 2017, Siegel 2018, Waseem and Hovy 2016, Zimmerman, Kruschwitz, and Fox 2018) and discouraging hateful speech (Munger 2017b, Munger 2017c).

Mass Partisan Polarization and Negative Partisanship

Although mass partisan polarization has been most commonly studied in terms of the divergence between Republicans' and Democrats' attitudes toward major policy issues (Abramowitz and Saunders 2008, Fiorina and Abrams 2008), more recent scholarship investigates affective polarization (or negative partisanship), the degree to which citizens dislike and distrust others identified with the other party (Iyengar et al. 2019). Pointing out the increasing affective polarization over the last several decades (Iyengar et al. 2019), the scholarship investigates its consequences, including anti-deliberative attitudes, social avoidance, and outright social discrimination (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, the cutting-edge research in U.S. politics takes one step further, demonstrating that a substantial minority do not just dislike out-partisans but also rationalize harm and even endorse outright violence against their partisan opponents (Kalmoe and Mason 2018).

Negative partisanship has mainly been measured using survey self-reports. While there exist a handful of other measurement approaches for affective polarization, such as IAT (Implicit Association Test) (Iyengar et al. 2019), survey self-reports have been the only strategy to measuring radical/lethal partisanship (Kalmoe and Mason 2018). My study is the first attempt to measure radical/lethal partisanship through its behavioral manifestation in online space. Although this approach shares with survey self-reporting 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. Besides, the current approach is a cost-effective tool for tracking the level of radical/lethal partisanship expressed on social media. With the approach, one can easily generate an uninterrupted data set by scraping and classifying social media posts.

Targeted Violent Political Rhetoric

I define violent political rhetoric as rhetoric expressing the intention of physical violence against political opponents, including outright threat, incitement, and endorsement. Here, the intention of physical violence 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 (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. Rather, essentially non-violent politics is figuratively described as events that involve physical violence such as a battle or a war.

Detecting Violent Political Rhetoric on Twitter

Detecting a highly specific subset of texts (e.g., violent political rhetoric) from a massive stream of social media posts poses a new challenge. This is because there is no pre-defined corpus from which to start classification. Although a small body of research on YouTube proposes several methods to identify threatening comments in YouTube videos on various controversial issues (Hammer et al. 2019, Wester 2016, Wester et al. 2016), they do not provide a systematic approach to defining an initial corpus that is comprehensive enough. In this section, I introduce a method that combines keyword extraction/filtering and machine learning to detect violent political rhetoric on Twitter (Figure 1).

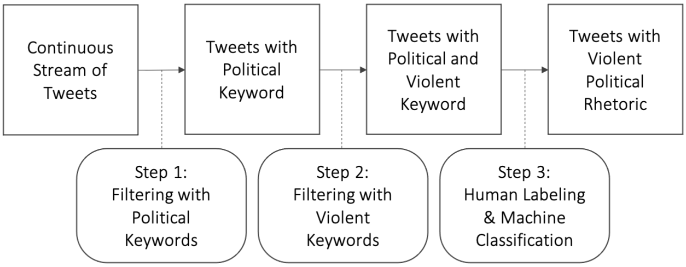

Continuous Stream of Tweets
Tweets with Political Keyword
Tweets with Political and Violent Keyword
Tweets with Violent Political Rhetoric
Step 1: Filtering with Political Keywords
Step 2: Filtering with Violent Keywords
Step 3: Human Labeling & Machine Classification

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).[1] 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 major politicians' Twitter accounts (members of Congress, governors, the President, and the Vice President, and the presidential and vice-presidential candidates) as well as those belonging to major parties.[2] The list of political keywords was compiled to serve two purposes. First, it is to collect political tweets. The tweets filtered and scraped through the list of the accounts are ones that "mention" at least one major political entity in U.S. politics. Naturally, the keywords of my choice indicate that the retrieved tweets are political in nature (or by definition). Second, the inclusion of a politician's account in a tweet containing a threat of violence can serve as an indicator that the politician is the target of the threat. That way, we can investigate who the targets of violent political rhetoric are and how they are distributed on party, gender, or the type of office.

I run a Python program that scrapes live tweets that contain any of the keywords in the

[1] The Twitter API is used to retrieve data and engage with the communication on Twitter.
[2] The full list of the political keywords can be found here.

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

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 existing corpus through those keywords. The only difference is that the filtering on violent keywords is conducted on a local machine, not on the Twitter API. 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 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 reveals the intention of violence or not. Here, I intend to extract violent keywords from a corpus that already contains information about what multiple people deem to be a threat of violence. Specifically, I use a corpus built by Jigsaw, a unit within Google (Jigsaw 2020).[3] The corpus contains around two-million Wikipedia comments labeled by multiple human coders for various toxic conversational attributes, including "threat." I fit a logistic regression classifier and extract terms (uni- and bi-gram features) that are most predictive of threatening intention 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 tweets from Step 1 into a violent political corpus and a non-violent political corpus.

Step 3: Human Labeling and Machine Classification

Although the previous round of filtering relies on a list of violent keywords that people frequently use in online space as well as consider to be violent, only a small fraction of the violent-keyword tweets actually contain the intention of violence. This is because many tweets contain a violent keyword without having any intention of violence in various ways. For instance, see how a violent keyword *shoot* is used in the following tweets: 1) "Biden told people there how police should *shoot* unarmed people, liberals are way too easily impressed." 2) "@realDonaldTrump Someone please *shoot* this b***h already." We can see that the keyword, *shoot*, is used to simply deliver information about Biden's remark in 1) while it is used to actually promote violence in 2).

To handle this, I manually-labeled a sample of tweets with political and violent keywords and built various machine learning classifiers. Specifically, I used active learning

[3] Click here for the corpus.

(Linder 2017, Miller, Linder, and Mebane 2020, Settles 2009). In active learning, we start with manually-labeling *randomly selected* texts, train a classifier, make predictions on unseen texts, *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. Note that the corpus compiled through Step 1 and 2 is highly imbalanced data with only a small fraction containing violent political rhetoric. In this case, 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 a tweet (see Appendix A and B for detailed description of the whole sequence of detecting violent political rhetoric).

Characteristics and Spread of Tweets Containing Violent Political Rhetoric

How prevalent is violent political rhetoric on Twitter? Does the rise of tweets containing such rhetoric respond to the offline world politics? What types of political figures appear and are targeted with such violent rhetoric? Who are the users who threaten and incite violence against political opponents? How diffusive is violent political rhetoric and what predicts its spread? In this section, I provide comprehensive data analysis concerning the characteristics and the spread of tweets containing violent political rhetoric. The following analysis is based on the data set collected between September 23 2020 and January 8 2021. This 16-week period covers major political events concerning the 2020 Presidential Election, including the Capitol Riot and the suspension of the former President Trump's Twitter account for his association with the riot.

The key findings involve the following. First of all, violent political rhetoric on Twitter is closely related to contentious events/issues in the offline-world politics, spiking to its highest level in the days preceding the Capitol Riot. Next, 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. Users who write violent tweets are ideologically extreme and located on the fringe of the Twitter network. While these users tend to be liberal, the ideological makeup varies over time depending on what issues violent political rhetoric arises from. In terms of targeting, violent tweets are disproportionately directed at women and Republican politicians. Finally, spread of violent tweets takes place primarily among ideologically similar users but there is also substantial amount of cross-ideological spread, raising concerns about co-radicalization.

Content of Violent and Non-violent Political Tweets

To shed light on how tweets containing violent political rhetoric differ from non-violent political tweets in terms of content, Figure 2 and Table 1 show 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 (Monroe, Colaresi, and Quinn 2008). In Figure 2, the x-axis indicates the relative frequency with which the keyword occurs in each type of tweets (violent vs. non-violent). The y-axis in each panel (non-violent on the top and violent at the bottom) depicts the extent to which the keyword is associated with each type of tweets. In the same vein, Table 1 lists the top-30 words by their type-specificity (words that are more frequently used for each type of tweets). Note that some of the words included as indicating violent political tweets have already been baked in as part of the violent-keyword filtering (see p. 12).

What is most noteworthy is that words that indicate certain political entities are much more frequent for violent tweets than for non-violent ones. In Table 1, we can see that, while no entity-specific terms were included in the top-30 list for non-violent tweets, the violent terms include many accounts that belong to high-profile political figures such as @realdonaldtrump (Donald Trump), @senatemajldr (Mitch McConnell), @mike_pence & @vp (Mike Pence), and @secpompeo (Mike Pompeo). In particular, the account for Trump, "@realdonaldtrump" is ranked third on the list, demonstrating 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., president, vote, ballot, state, tax, county, electoral, campaign) along with words that represent particular political events/issues such as "georgia" (the Senate election in Georgia) or "fraud" (misinformation about election fraud).

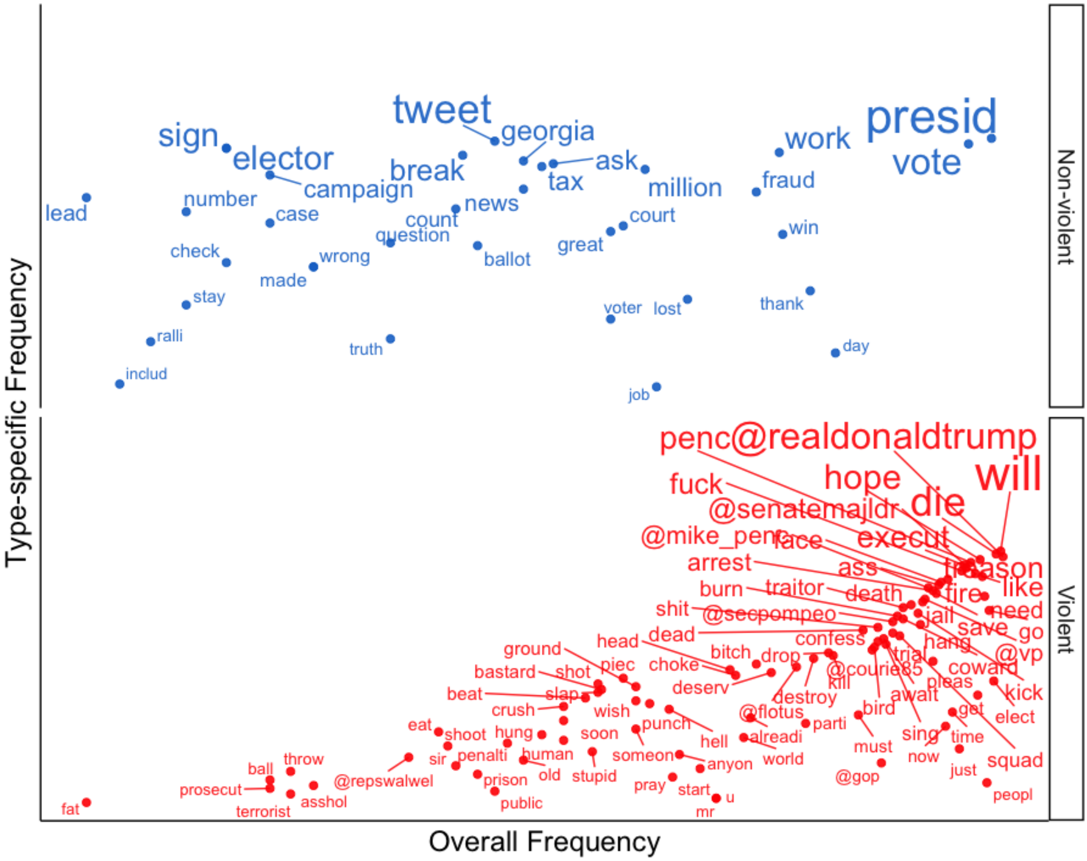


Figure 2. *Comparison of terms by type of tweets*

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.

TABLE 1 *Comparison of terms by type of tweets*

Rank	Non-violent	Violent	Rank	Non-violent	Violent
1	presid	will	16	answer	save
2	tweet	die	17	investig	need
3	vote	@realdonaldtrump	18	counti	@vp
4	sign	hope	19	news	jail
5	elector	penc	20	fraud	death
6	work	execut	21	pa	traitor
7	break	treason	22	lead	go
8	trust	fuck	23	pennsylvania	kick
9	georgia	@senatemajldr	24	video	burn
10	ask	like	25	report	coward
11	tax	fire	26	count	@secpompeo
12	million	face	27	number	hang
13	lt	ass	28	congratul	shit
14	campaign	@mike_penc	29	seem	dead
15	retweet	arrest	30	communiti	trial

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 2 reports the top-30 hashtags that are most frequently used in tweets containing violent political rhetoric. 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 (#savebrandonbernard,

#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 each and every hashtag in the list and the corresponding issues (some will be discussed in the next section), they together make it clear that violent political rhetoric responds to or arises from various political/social issues in offline politics.

TABLE 2 *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

Timeline of Violent Political Tweets

Then, how frequent are tweets containing violent political rhetoric? Figure 3 illustrates the timeline of tweets containing violent political rhetoric. The trend is expressed in terms of the count and of the 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, Nikitin, et al. (2019) 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 that threaten and incite violence, 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.[4]

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

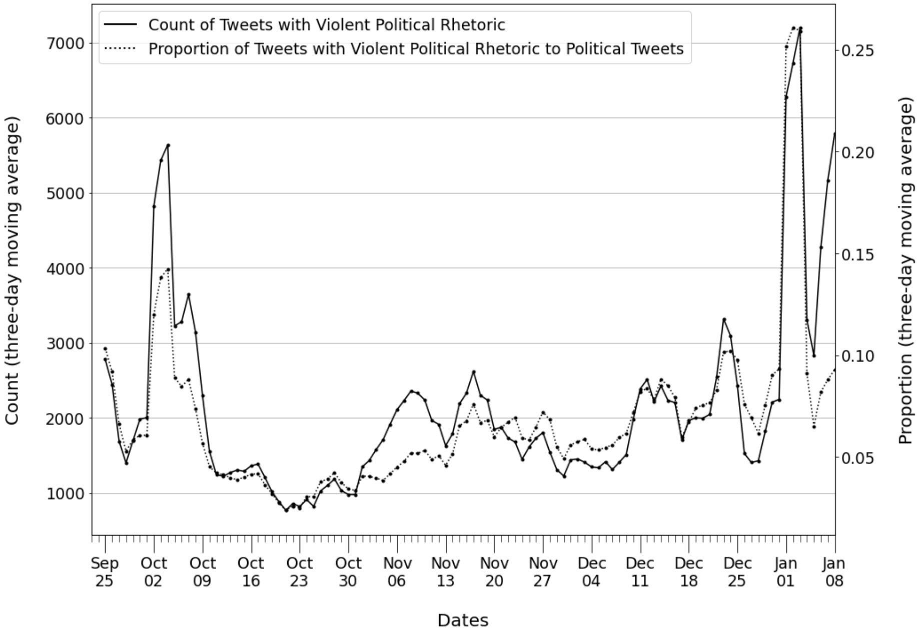


Figure 3. *Timeline of violent political rhetoric (September 23 2020 - January 8 2021)*

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.

Then, what issues give rise to the outburst of violent political rhetoric? As illustrated in Figure 3, there is a considerable over-time variation in the trend of 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 day (November 4 2020). To get a sense of what issues drive the trend, Table 3 reports the weekly top-5 hashtags included in violent tweets. While the steady uptrend toward the election day appears associated with the partisan competition/tension over the election (#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 violent intention against Trump, including many death wishes.

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 violent rhetoric such as "#arrest and #execute #traitortrump", together 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 (e.g., the death penalty in the week of December 9 - 15), 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,[5] it

[5] 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.

is abundantly clear that offline political conflict is intertwined (potentially causally) with violent political rhetoric on Twitter.

TABLE 3 *Most frequent hashtags in violent political rhetoric (weekly)*

1	(2020) 9/23-9/29	9/30-10/6	10/7-10/13	10/14-10/20
2	#trump2020	#covid19	#executed	#treason
3	#maga	#vote	#amendments	#biden
4	#treason	#walterreed	#bancapitalisim	#sealteam6
5	#debates2020	#trump	#constitution	#hillaryclinton
6	#whenthesecondwavehits	#covidiot	#government	#obama
7	10/21-10/27	10/29-11/3	11/4-11/10	11/11-11/17
8	#crimesagainstchildren	#endnigeria	#jesuschrist	#antifaarefascists
9	#crimesagainsthumanity	#endsars	#trump2020	#blmareracists
10	#laptopfromhell	#vote	#trump	#marchfortrump
11	#tonybobulinski	#trump2020	#maga	#trump2020
12	#moscowmitch	#electionday	#trumpcrimefamily	#treason
13	11/18-11/24	11/25-12/1	12/2-12/8	12/9-12/15
14	#treason	#maga	#treason	#savebrandonbernard
15	#maga	#diaperdon	#magabusmusts	#brandonbernard
16	#scif	#fightbacknow	#magaqueentrains	#gopisover
17	#trump	#richardmoore	#bidencheated2020	#abolishthedeathpenalty
18	#democracydemandsit	#headsmustroll	#kag2020	#treason
19	12/16-12/22	12/23-12/29	12/30-1/5 (2021)	1/6-1/8
20	#pardonsnowden	#wethepeople	#fightback	#maga
21	#freeassange	#1	#1776again	#traitortrump
22	#punkaf	#pencecard	#godwins	#execute
23	#wethepeople	#pardonsnowden	#divinetiming	#arrest
24	#stopthesteal	#freeassange	#trustgod	#trampicantraitors

One venue for future investigation with regard to the over-time variation is to look at how offline political violence affects violent political rhetoric in online space (or vice versa). Figure 4 illustrates the timelines of violent political rhetoric on Twitter and of offline political violence. Although this comparison based on aggregate measurement does not provide conclusive evidence for the relationship between the two, the general trend exhibits certain similarities, particularly after the election. Given the increasing inter-party animosity expressed in the form of offline political violence and the danger it poses to democracy, it is crucial for future research to investigate whether and how offline political violence translates into violent political rhetoric in online political communication (or vice versa).[6]

[6] The cutting-edge research in political communication provides abundant evidence that online and offline political events are closely related although the causal direction between the two remains to be explained (Chan, Ghose, and Seamans 2016, Gallacher, Heerdink, and Hewstone 2021, Klein 2019, Mooijman et al. 2018). With particular relevance to the current finding, a stream of research examines how offline political events, including protests, violence, and rise of certain politicians affect aggressive political rhetoric in online space (Olteanu et al. 2018, Siegel, Nikitin, et al. 2019, Vegt et al. 2019, Wei 2019).

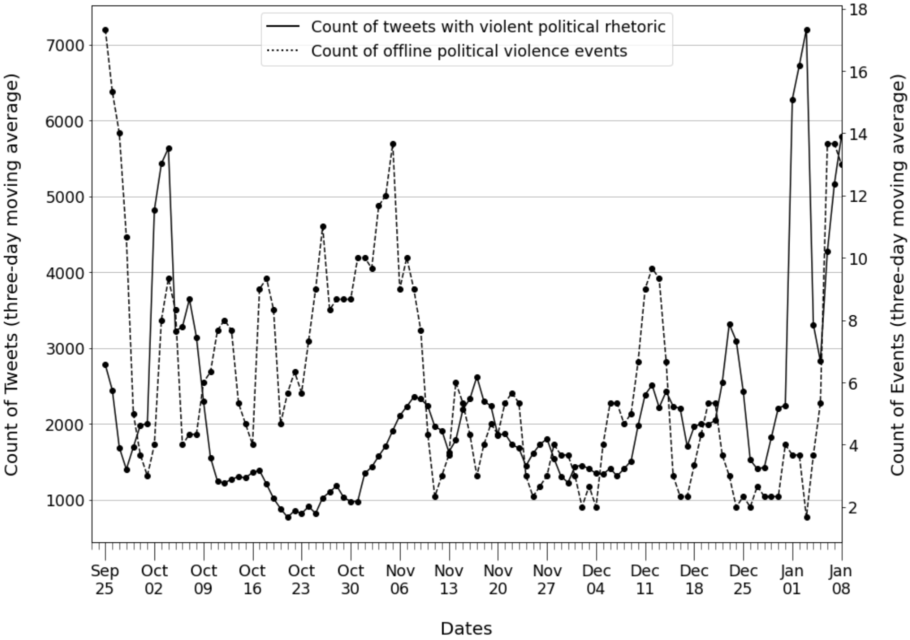


Figure 4. *Timeline of violent political rhetoric and offline political violence (September 23 2020 - January 8 2021)*

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 number of offline political violence events. Each point in the lines indicates the three-day moving average. The data for offline political violence is from U.S. Crisis Monitor (USCM) led by Armed Conflict Location and Event Data (ACLED) and the Bridging Divides Initiative (BDI). USCM provides real-time data and on political violence in the U.S. For the count of violent political events, I include the following types of events in violent political events (from the “SUB_EVENT_TYPE" variable in the data): violent demonstration, protest with intervention, excessive force against protesters, attack, mob violence, arrests, looting/property destruction, armed clash, disrupted weapons use, sexual violence, suicide bomb, and grenade. For more detailed information on the data and coding rules, visit U.S. Crisis Monitor.

Political Accounts in Violent Tweets

As previously discussed, in the era of radical/lethal partisanship, a substantial minority of citizens see the use of violence against political opponents as acceptable (Kalmoe and Mason 2018). Indeed, the Capitol Riot and a number of ordinary Republicans' approval of the riot (Spoccia 2021) are an embodiment of such extreme partisanship. Then, what politicians appear rhetoric and are threatened in violent tweets?

Accounts can be "mentioned" either by including the account in a tweet or by replying to politicians' tweets. Tables 4, 5, and 6 report what accounts are mentioned in violent tweets and present them by the type of office, partisan affiliation, and gender. Each table records the average number of violent tweets that mention political account(s) in a given category.

First of all, Table 4 shows that Trump is at the center of violent partisan expression 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 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 expression.

Given that Trump (Republican and man) can obscure the comparison based on partisan affiliation and gender, Tables 5 and 6 report the relevant statistics without violent tweets that mention his account. Table 5 shows that Republicans appear more frequently than non-Republicans (Democrats and a handful of independent/minor-party politicians). In Table 6, we can see that, on average, men politicians appear more frequently in violent tweets than women politicians.

TABLE 4 *Mean mention count by office type*

	Mention Count
Trump	137,475
Pence	18,506
Biden	8,759
Harris	467
Governors	165
Senators	478
Representatives	56

TABLE 5 *Mean mention count by partisan affiliation*

	Mention Count
Republican	257
Non-Republican	117

TABLE 6 *Mean mention count by gender*

	Mention Count
Women	103
Men	207

To further explore how the type of office, partisan affiliation, and gender correlate with the mentioning/targeting of political accounts in violent tweets,[7] Table 7 reports the

[7] Although mentioning is often used to measure targeting in the literature (Munger 2017a, Siegel, Nikitin, et al. 2019), not all mentions indicate targeting. To get a sense of 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

results from a negative binomial regression analysis where the count of mentions in violent tweets, the response variable, is regressed against the type of office, partisan affiliation, 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 account (or visibility). In order to prevent a tiny subset of the observations from being overly influential, I exclude the candidates for the presidential election (Biden, Trump, Harris, Pence). For the details of modeling and robustness analysis, see Appendix C.

First, the results reveal that being Republican correlates positively with mentioning/targeting in violent tweets (Model 5). Why do Republican politicians appear and are targeted 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/targeting in violent tweets (Model 5). This is consistent with evidence, both academic and journalistic, for online aggression that disproportionately targets women politicians (Cohen 2021, Di Meco and Brechenmacher 2021, Felmlee,

politician. The coding rules are 1) the tweet clearly expresses the intention of violence, 2) the intention of violence in the tweet is clearly targeted at a specific politician (one of the political accounts included for regression analysis), 3) the tweet mentions the account of the targeted politician. The result shows that about 40% of violent tweets target the politician whose account is mentioned in the text. This proportion is slightly higher than what is found in a similar study on hate speech in online political communication (about 25% in Siegel, Nikitin, et al. 2019).

TABLE 7 *Mentioning/targeting of political accounts: negative binomial regression*

	Model 1	Model 2	Model 3	Model 4	Model 5
Office:governor	1.08^{***}				0.51^{*}
	(0.30)				(0.22)
Office: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)				1.12^{***}	1.09^{***}
				(0.05)	(0.05)
(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.

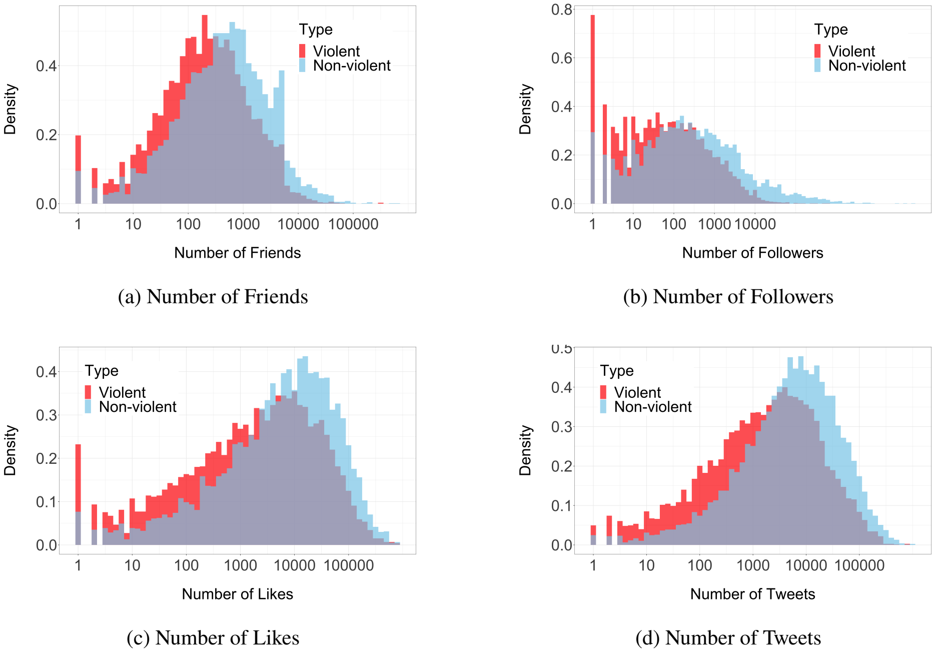
Rodis, and Zhang 2020, Fuchs and SchÄfer 2019, Jones 2020, Rheault, Rayment, and Musulan 2019, Southern and Harmer 2019).

Engagement in Political Communication Network by Tweeter Type

How active are users who use violent language in the political communication network on Twitter? This question is important because the more central to the network violent tweeters are, the more likely ordinary users are connected to violent tweeters and exposed to violent political rhetoric. Table 8 and Figure 5 report the summary statistics and logged distribution for user-level indicators of engagement in the political communication network on Twitter. Together, they show that violent users are less active and their personal networks are more sparse than non-violent users'. Violent users connect to other users, "like" others' tweets, and publish their own tweets to a lesser degree than non-violent users. In particular, the smaller number of followers and tweets imply that exposure to violent tweets can be generally limited than non-violent political users' content. In sum, violent users are located on the fringe of the political communication network on Twitter.

TABLE 8 *Median Value for network engagement indicators*

Count	Violent	Non-violent
Friends	205	425
Followers	53	193
Likes	2,275	6,663
Tweets	1,841	5,784



(a) Number of Friends

(b) Number of Followers

(c) Number of Likes

(d) Number of Tweets

Figure 5. *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 Figure 6, 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). 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 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 4). 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 (Figures 7). Although the difference decreases 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. Figure 8 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 3 — 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. Figure 9 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 in Figure 10. 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.

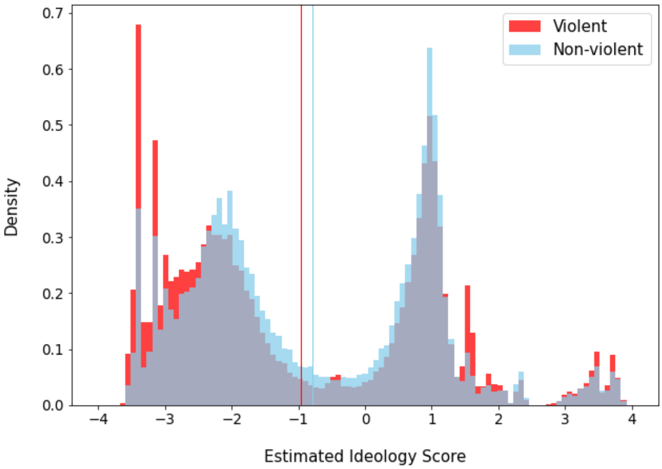


Figure 6. *Distribution of ideology by type of political tweeters*

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.

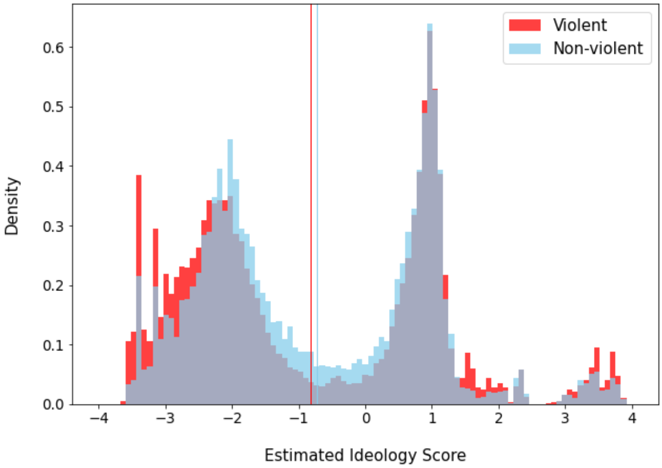


Figure 7. Distribution of ideology by type of political tweeters (without Tweets mentioning '@realDonaldTrump')

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.

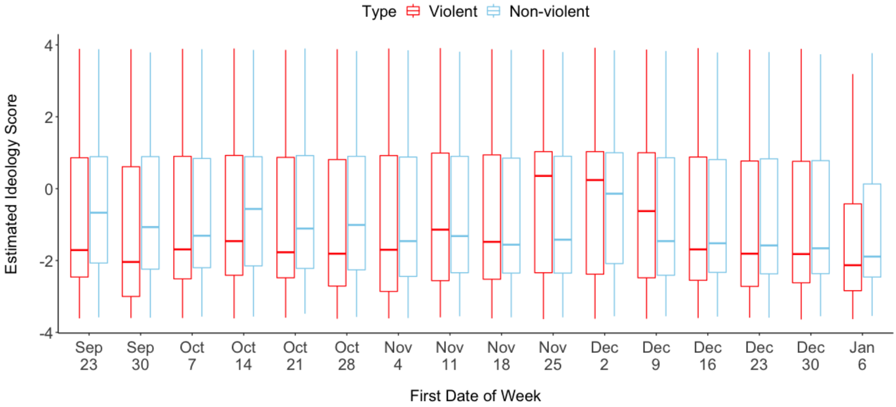


Figure 8. Weekly distribution of ideology by type of political tweeters

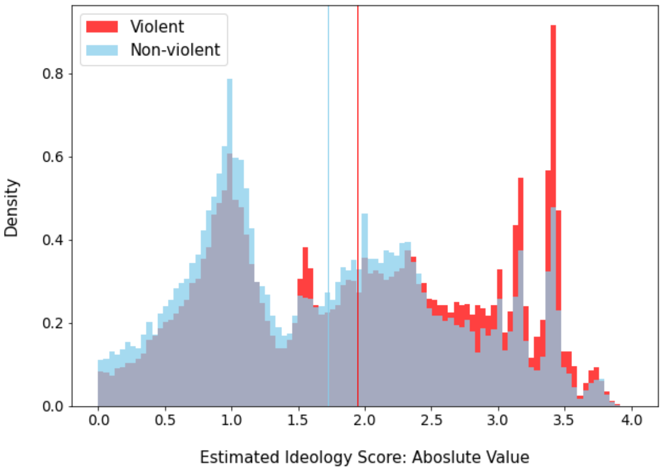


Figure 9. *Distribution of ideological extremity by type of political tweeters*

Note: The unit of observation is a user. The *x*-axis depicts the ideological extremity score with larger values indicating more extremity. The *y*-axis is probability density. The vertical lines indicate the mean value for each group.

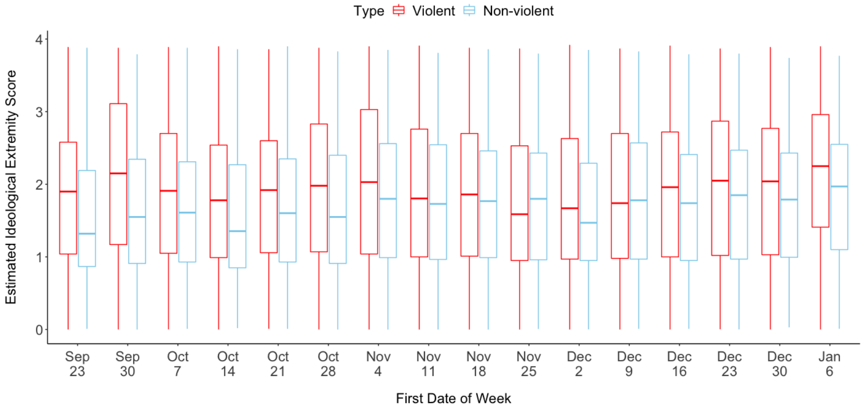


Figure 10. *Weekly distribution of ideological extremity by type of political tweeters*

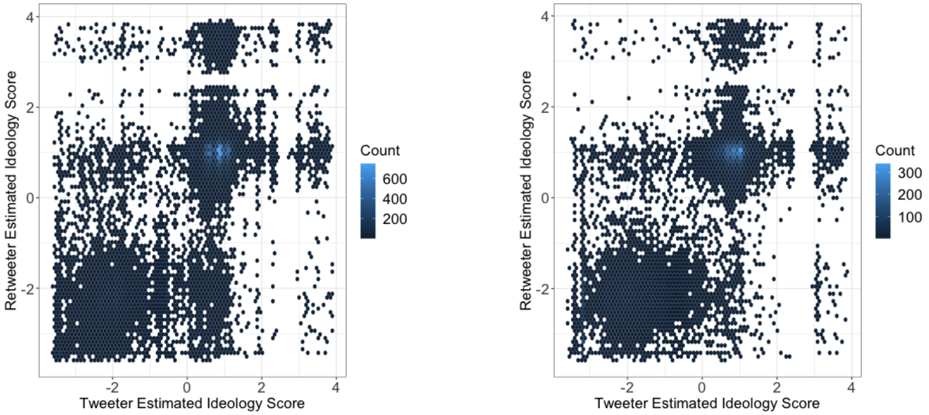
Spread of Violent Political Rhetoric

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? Figure 11 presents 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 political retweeting, both violent and non-violent, is affected by ideological homophily (a tendency for individuals to form ties with those who are ideologically close to themselves), 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). It implies that, while liberal and conservative violent users tend to flock together with the liked-minded, they can also encounter and spread partisan opponents' violent behavior, potentially co-radicalizing each other by feeding off partisan opponents' violent intention (Ebner 2017, Knott, Lee, and Copeland 2018, Moghaddam 2018, Pratt 2015).

Then, how far do violent tweets travel on the Twitter communication network? 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. Figure 12 describes the distribution of the shortest path distance on the following network for all the retweets 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 1 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 2 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 the condition is met, the shortest distance is estimated as 3 or more.

As seen in the figure, almost two-thirds of the retweets take place between a pair of users with a direct tie (62%). However, there is a substantial minority of retweets that travel beyond the tweeter's followers. Around 31% of the retweets take place between users whose estimated shortest path distance is 2. For the remaining 7%, tweets were retweeted over three or more ties. It means that even if users do not follow a violent tweeter (even the follower of a violent tweeter), it is still possible for them to get exposed to such discomforting content against one's intent. Besides, 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.



(a) Violent tweets

(b) Non-violent tweets

Figure 11. *Ideological homophily: correlation between tweeters' and retweeters' ideology*

Note: 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.

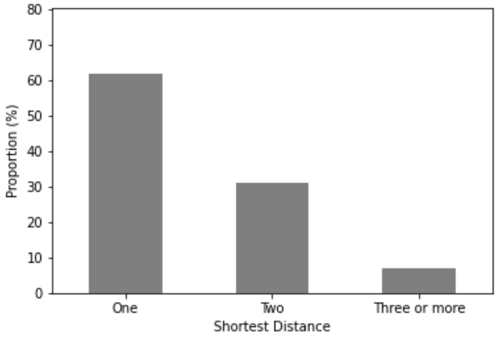


Figure 12. *Reach of tweets containing violent political rhetoric on the following network*

Note: 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.

Conclusion

The recent violent hostility among ordinary American partisans, as dramatically expressed in the Capital Riot on January 6 2021, has drawn immense attention both from the media and from 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 in online space and thus contribute to the fields of grassroots political violence, online political communication, and radical/lethal 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 radical/lethal 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 in the offline world. Also, users who threaten violence are ideologically extreme and located on the fringe of the communication network. In terms of targeting, violent threats are disproportionately 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-mind 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 2015).

The findings in this paper call for future research on violent political rhetoric as well. First, what are the causal relationships between violent political rhetoric in online space and offline political violence? While this paper presents abundant evidence for close relationships between the two (see also Gallacher, Heerdink, and Hewstone 2021), 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, Kalmoe 2014, Kalmoe, Gubler, and Wood 2018, Kalmoe 2019), little attention has been paid to an extreme form of violent language such as a threat of violence. Therefore, it is crucial to investigate the effects of exposure to an explicit threat for 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?

Appendix

A. Manual Annotation of Violent Political Rhetoric

Three human coders (including the author and two undergraduate students in the Department of Political Science at Penn State) labeled tweets in terms of whether a given tweet expresses the intention of political violence or not. 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 coders were asked whether a given tweet's author expresses any intention of violence (including a threat, endorsement, incitement of physical violence against a political entity) or not.

The concept of the intention of violence is inherently ambiguous and subjective. Therefore, it was necessary to refine the conceptualization and operationalization by developing a detailed coding guideline throughout the manual annotation process. The major sources of false positives involve 1) when violent phrases are used as a metaphor that describes non-violent political events as violent (Kalmoe 2013, Kalmoe 2014, Kalmoe, Gubler, and Wood 2018, Kalmoe 2019), 2) a religious curse that does not involve any real violence (e.g., 'burn in hell!'), 3) mentioning (or even criticizing) violent intention expressed by someone else, and 4) sarcasm (e.g., 'why don't just shoot them all if you believe violence solves the problem?'). Detailed coding rules for dealing with such confusing cases are documented in this Github repository.

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 7,500 tweets, three coders worked on three different sets of tweets (Coder 1: 3,500, Coder 2: 3,500, Coder 3: 500). In sum, a total of 10,000 tweets were manually labeled.

As previously noted, the concept of the intention of violence is inherently vague and subjective. Accordingly, the levels of inter-coder agreement for similar studies on aggressive online behavior are generally moderate (Table A1). The inter-coder agreement scores achieved in our study are reported in Table A2. 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 A1 *Inter-coder agreement on similar concepts*

Study	Concept	Krippendorff's Alpha
Theocharis et al. (2016)	political incivility	0.54
Munger (2017a)	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 A2 *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		0.597	
Fless's Kappa		0.597	
Krippendorff's Alpha		0.597	

B. Active Learning and Classification

Relying on active learning (Linder 2017, Miller, Linder, and Mebane 2020, Settles 2009), I followed the next process to build a 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 another 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 another 500 tweets for additional manual annotation (for detailed information about BERT, see Devlin et al. 2018). Through this iterative process, a total of 10,000 tweets containing political and violent keywords are manually labeled.

With the final training set of size 10,000, I fit a series of machine learning classifiers. Since the data set is imbalanced, I used precision, recall, and F-1 score to evaluate the performance of classifiers. The results are reported in Table A3. As shown in the Table, the BERT model achieves the best performance and is used for final classification. For the BERT model, the binary decision threshold is set at 0.925 since most relevant cases start to appear on the right tail of the probability distribution.

Note that the 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 the European context. Using the data set, Wester (2016) and Wester et al. (2016) build a series of statistical classifiers with various lexical and linguistic features and find that the best performance was achieved by 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).

TABLE A3 *The average performance of classifiers from 5-fold cross validation*

Model	Precision	Recall	F-1
Logistic Regression + Count Vector	68.64	32.78	44.33
Logistic Regression + TF-IDF Vector	80.75	9.91	17.62
Logistic Regression + GloVe	60.58	10.66	18.09
Random Forest + Count	77.83	19.17	30.67
Random Forest + TF-IDF Vector	80.69	17.34	28.50
Random Forest + GloVe	74.14	10.97	19.02
XGBoost + Count Vector	78.15	7.74	14.06
XGBoost + TF-IDF Vector	79.49	11.67	20.28
XGBoost + GloVe	68.15	14.18	23.46
BERT	74.02	59.05	65.65

C. Regression Analysis on Mentioning/Targeting

Table A4 reports descriptive statistics for the mentioning/targeting analysis. Tables A5 and A6 report two additional models to assess whether the findings in Table 7 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 family is used for all of the models to deal with over-dispersion. As seen in the coefficients, the results for office type, gender, and partisan affiliation are consistent across the models.

TABLE A4 *Descriptive statistics for mentioning/targeting analysis*

Mention Count	Folloew Count	Gender	Party	Office
Min. : 0.0	Min. : 2,496	Women: 136	D :303	Representative: 436
1st Qu.: 2.0	1st Qu.: 21,772	Men: 449	DFL: 1	Governor: 50
Median : 6.0	Median : 37,047		I : 2	Senator: 99
Mean : 136.7	Mean : 191,013		L : 1	
3rd Qu.: 27.0	3rd Qu.: 105,734		R :278	
Max. :25266.0	Max. :12,102,376			

TABLE A5 *Mentioning/targeting of political accounts: negative binomial regression + Biden/Pence/Harris*

	Coefficient (S.E.)
Office:Biden	−0.30
	(1.44)
Office:Pence	1.19
	(1.42)
Office:Harris	−2.96*
	(1.42)
Office:governor	0.51*
	(0.22)
Office:senator	0.18
	(0.18)
Women	0.97***
	(0.15)
Republican	0.99***
	(0.13)
Follower Count (log)	1.09***
	(0.05)
(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 A6 *Mentioning/targeting of political accounts: zero-inflated negative binomial regression*

	Coefficient (S.E.)
Count model: (Intercept)	-9.44^{***}
	(0.56)
Count model: Office:governor	0.49^{*}
	(0.21)
Count model: Office: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: Office:governor	−16.07
	(3332.84)
Zero model: Office: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

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

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.

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.

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.

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.”

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, 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.

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.

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 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.

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.

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.

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.

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.

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."

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.*

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.*

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.

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. 2017a. “Don’t@ Me: Experimentally Reducing Partisan Incivility on Twitter.” *Unpublished manuscript.*

———. 2017b. “Experimentally Reducing Partisan Incivility on Twitter.” *Unpublished working paper. Available at: https://kmunger. github. io/pdfs/jmp. pdf.*

———. 2017c. “Tweetment effects on the tweeted: Experimentally reducing racist harassment.” *Political Behavior* 39 (3): 629–649.

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. 2015. "Islamophobia as reactive co-radicalization." *Islam and Christian–Muslim Relations* 26 (2): 205–218.

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, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2019. “Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath.”

Siegel, Alexandra, Joshua Tucker, Jonathan Nagler, and Richard Bonneau. 2019. “Tweeting beyond tahrir: Ideological diversity and political tolerance in egyptian twitter networks.” *Unpublished working paper, New York University.*

Siegel, Alexandra A. 2018. “Online Hate Speech.”

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.

Spoccia, Gino. 2021. “45% of Republicans approve of the Capitol riots, poll claims.” *Independent* (July). https://www.independent.co.uk/news/world/americas/us-election-2020/republicans-congress-capitol-support-trump-b1783807.html?fbclid=IwAR0HjPkesWHmINvYr1jLwlVbgC2h3_uK-PulRTF8R8DVnXlxD_TUHVk-E_U.

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

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.

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, 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. “Filter realtime Tweets.” Accessed February 17, 2021. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview.

———. 2021b. “Search Tweets: standard v1.1.” Accessed February 17, 2021. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets.

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.

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.

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.

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).*