<Working Paper>
Please do not cite. If you have a question, email me at taegyoon@psu.edu

Violent Political Rhetoric on Twitter *

TAEGYOON KIM

Recently, concerns are raised about the violent nature of political communication on social media: many social media platforms are exploited by users who write posts threatening violence against political opponents as part of normalized expression of radical partisanship. Focusing on Twitter, I develop a method for automatic detection of violent political rhetoric. I apply the method to collect a data set of Tweets containing violent political rhetoric, spanning the 14-week period surrounding the 2020 Presidential Election. Using the data set, I investigate the characteristics and the spread of violent political rhetoric on Twitter. The key findings include a) Violent political rhetoric is rare (0.07% of political communication) but is spread through multiple chains on following ties (about 10% of retweets traveling over at least three ties), incidentally exposing a potentially huge size of audience and thereby amplifying its effects. b) users who write violent Tweets are on the fringe of the Twitter network and are ideologically more extreme and liberal than non-violent users. c) Spread of violent Tweets takes place primarily among ideologically-homogeneous users but there is also substantial amount of cross-ideological exposure.

The emergence of social media platforms was widely touted as a technological revolution that would bring about many beneficial outcomes such as political learning and participation (Dimitrova et al. 2014, Tucker et al. 2017). However, such early hopes are being overshadowed by mounting concerns about aggressive political communication. Nowadays, it is not difficult to find uncivil political discussion both from political elites and from ordinary users. Also, various types of hate speech — targeted at females, 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 languages and curb their spread (Siegel 2018). However, we know very little about another, perhaps most deleterious, type of aggressive political communication: violent political rhetoric. Violent political rhetoric, expressing intention of violence against political opponents in its extreme form, has drawn significant media attention. Numerous media reports show

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

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, Vigdor 2019, Daugherty 2019). In particular, many social media platforms were implicated in the extremists' effort to motivate and organize the Jan 6 Storming of the Capitol that left a vivid and deep scar on the U.S. democracy. Importantly, there is plenty of evidence that it is on both niche extremist online forums and mainstream social media platforms, including Twitter, where posts expressing intention to storm the Capitol surged towards Jan 6 2021 (Guynn 2021, Lytvynenko and Hensley-Clancy 2021, Romm 2021).

Violent political rhetoric is worrisome not only because it serves as a signal of extremist offline violence but also because exposure to such rhetoric has negative political 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 be widely spread through the communication network on social media, amplifying its negative effects. In addition, such rhetoric is in itself a behavioral manifestation of radical/lethal partisanship (Kalmoe and Mason 2018) where individuals not just hate out-partisans (Abramowitz and Webster 2018) but also support and even enjoy the use of offline violence against them. This is an online mirror image of the recent instances of violent clash between Republicans and Democrats revolving around major political issues, including Black Lives Matter movements as well as the appalling violent incidents in the wake of 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 the offline world politics? How diffusive is it and what predicts its spread? What types of political figures are targeted with such violent rhetoric? Who are the users who threaten violence against political opponents? Given the significance of violent political rhetoric, it is imperative to investigate these questions but little scholarly effort has been spent on this. Due to the massive size of 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 reports (Lytvynenko and Hensley-Clancy 2021, Romm 2021). Therefore, I develop an automated method based on various computational tools to detect violent political posts from continuous streams of data, focusing on Twitter. I then apply the method to build a data set of Tweets containing violent political rhetoric over the 14-week period surrounding the 2020 Presidential Election. Finally, I provide comprehensive data analyses 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 space. While a body of research in offline political violence has taken a bottom-up approach to study why individuals take part in collective violence in the offline world (Claassen 2016, Horowitz 1985, Fujii 2011,

Scacco 2010, Tausch et al. 2011), few studies have taken a bottom-up approach to explore why individuals threaten violence against political opponents in online space and who the targets of such threats are. I fill part of the gap by showing that individuals who write Tweets containing a threat of violence against political opponents are ideologically more extreme than ordinary individuals and that political elites are often targeted but to a varying degree by political position, gender, and partisan affiliation. Further, I open up a new research opportunity for the relationship between online threats of political violence and physical violence in the offline world.

By identifying and characterizing violent political rhetoric on Twitter, I also extend the study of aggressive political communication in online space where incivility and hate speech have been the key areas of inquiry. The literature in political communication shows that various forms of aggressive political speech can have heterogeneous effects (Zeitzoff 2020) and violent political rhetoric, in particular, can even lead to support for physical political violence (Kalmoe 2014). However, lack of a scalable method for data collection have hindered researchers from grasping the characteristics of violent political rhetoric in online space. My analysis shows that, similar to other aggressive political speech (Siegel, Nikitin, et al. 2019)), Tweets containing violent political rhetoric are rare on Twitter (0.05% of political Tweets, on average) and Tweeters who use such rhetoric tend to lie on the fringe of the political communication network than non-violent users. At the same time, however, it also shows that a substantial proportion of violent Tweets spread beyond the direct ties of their original author: 35% of the Retweets of such Tweets spread through indirect ties (i.e., my friend's friend, a friend of my friend's friend, etc), there by creating huge potential for incidental exposure to counter-attitudinal information. These findings extend the scope of research on aggressive online political communication that was largely focused on incivility and hate speech (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).

Finally, I shed light on the literature on partisan polarization and negative partisanship. Recent studies on polarization have highlighted that partisans are not just ideologically far apart (Abramowitz and Saunders 2008, Fiorina and Abrams 2008) but dislike or endorse physical violence against each other (Abramowitz and Webster 2018, Iyengar, Sood, and Lelkes 2012, Iyengar et al. 2019, Kalmoe and Mason 2018). However, there is little effort to investigate how such extreme forms of partisanship are expressed in online space. My work contributes to this literature by demonstrating radical/lethal partisanship revealed in the form of writing Tweets violent political rhetoric against out-partisans. The data analysis further highlights the ideological characteristics of violent Tweeters and the spread of violent Tweets. With regard to ideology, violent Tweeters are ideologically more extreme than ordinary Tweeters. The ideological distribution among is fairly symmetric between liberals and conservatives. In terms of spread, retweeting of violent Tweets takes place primarily among ideologically similar individuals — but not particularly to a greater extent than ordinary political communication — while there is still a considerable amount

of cross-ideological exposure.

Related Work

Here, I introduce three strands of literature, each of which provides context for as well as theoretical insight into the current study on violent political rhetoric in online space. a) Not only does my study expand the literature on individuals' participation in offline political violence but it can also benefit from the literature: 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. b) Although my work is among the first to investigate threats of political rhetoric in online space, an extensive body of research investigates political elites' use of violent metaphors in offline context and incivility and hate speech in online political discussion, providing rich context for the inquiry into violent political rhetoric in online space. c) A threat of political violence is a form of behavioral manifestation of extreme negative partisanship. Thus, the literature on partisan polarization and negative partisanship is extremely useful in understanding why Twitter users express violent intent against the partisan opposition and the consequences of such behavior.

Offline Political Violence

Although no research exists 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 context, a group of works seek 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 (Humphreys and Weinstein 2008, Popkin 1979, DiPasquale and Glaeser 1998, Lichbach 1995), social pressure from in-group members (Fujii 2011, Scacco 2010, Fuji 2009, Scott 1976, Taylor 1988), and perceived inequality in the distribution of resources among groups (e.g. jobs) 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 highlight the role of politicians' rhetoric. Kalmoe (2014) finds that exposure to mildly violent political metaphors during electoral campaigns 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 inter-group violence. Focusing on the 2015 Baltimore protests, Mooijman et al. (2018) shows that moralization of political issues can lead to 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 potential consequences for various political outcomes. Focusing on violent political metaphors that describe politics as violent events such as a battle or a war, Kalmoe (2014) demonstrates that exposure to such rhetoric increases individuals' support for violence against political elites in the opposition. Other works investigate the relationship between violent political metaphors and non-violent political phenomena. 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 primes aggression in aggressive partisans and thus lead to intransigence on issue positions.

While violent political rhetoric is studied in 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). In addition, they report evidence that such content is becoming more and more prevalent on social media. Such aggressive online speech is reported to discourage participation in online discussion (Henson, Reyns, and Fisher 2013), exacerbates 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 such aggressive language (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 what is known as 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 has raised concerns by highlighting its negative consequences, including anti-deliberative attitudes, social avoidance, or outright social discrimination, (Hutchens, Hmielowski, and Beam 2019, MacKuen et al. 2010, Abramowitz and Webster 2016; Cho, Gimpel, and Hui 2013; Iyengar, Sood, and Lelkes 2012; Huber and Malhotra 2017; Klofstad, McDermott, and Hatemi 2012; Lelkes, Sood, and Iyengar 2017; Levendusky 2013; Mason 2015; Mason 2018; McConnell et al. 2018; Miller and Conover 2015; Pew 2014; Pew 2016a: Pew 2016b). Extending the study of negative partisanship, the cutting-edge research in U.S. politics takes one step further, demonstrating that a substantive 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 been mainly measured with self-reports from surveys. While there exist a handful of other measurement approaches for affective polarization such as IAT (Implicit Association Test) (Iyengar et al. 2019), survey self-reporting has been the only strategy to measuring lethal partisanship (Kalmoe and Mason 2018). My study is the first attempt to measure 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. In addition, my approach provides a cost-effective tool to measure the level of lethal partisanship expressed on social media: 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 political speech expressing intention of violence against political opponents. Here, it is important to note that political speech bears intention of physical violence against a political entity. Therefore, violent political rhetoric, by definition, is targeted. The target is typically the political opponent, either a group (e.g., Republican representatives, Democratic senators) or an individual politicians (e.g., Donald Trump, Joe Biden). Violent political rhetoric can be an outright threat, endorsement, or incitement of physical violence.

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 call for violence at the upper end." Closely related to my study is a type of the violent political rhetoric at the upper end of the spectrum: threats of political violence against political opponents.

Kalmoe and his coauthors' works focus specifically on violent political metaphors (Kalmoe 2014, Kalmoe, Gubler, and Wood 2018, Kalmoe 2019). In his work, violent political metaphors are defined as "figure 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 very specific subset of texts (like, violent political rhetoric) from a massive stream of social media posts pose a new challenge. This is because there is no pre-defined corpus from which to start a classification task. Although a small body of research on YouTube proposes a series of methods to identify threatening comments in YouTube videos on various religious/cultural/political 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.[1] In this section, I introduce a method that combines keyword extraction/filtering and machine learning (active learning) to detect violent political rhetoric on Twitter in real-time.

[1]Instead, they use human-selected nineteen YouTube videos which they deemed created fierce controversies.

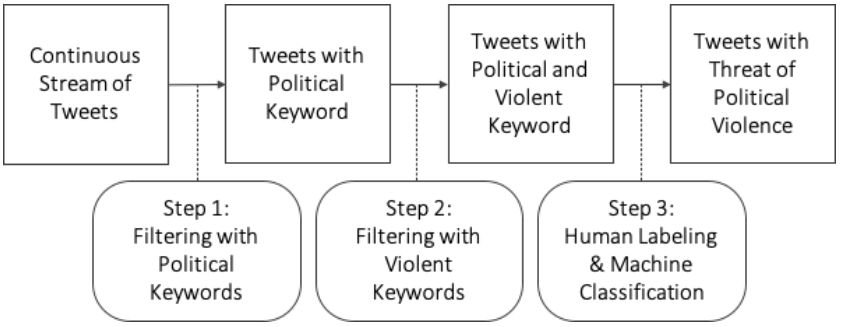


Figure 1. Data Collection Pipeline

Step 1: Filtering through Political Keywords

I start with compiling a list of political keywords to get Tweets from the Twitter API (Application Programming Interface).[2] Since a massive number of heterogeneous Tweets are generated in real time on Twitter, I first filter the Tweet stream through a set of political keywords. The keywords involve major politicians' Twitter accounts (members of Congress, Governors, President, and Vice President, etc) as well as those belonging to major parties.[3] The political keywords were compiled to serve two purposes. For one, 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 the U.S. politics. Naturally, the keywords of my choice highly likely indicate that retrieved Tweets are political in nature (or by definition). Also, 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 along party, gender, or race.

I run a Python program that scrapes live Tweets that contain any of the keywords in the list. The program is designed to scrape live Tweets continuously via the Twitter Streaming API (Twitter 2021a). This API allows researchers to scrape streams 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 potential for the platform to engage in censorship. That is, Tweets retrieved via Search API can leave out violent political rhetoric because Tweets involving it might have been deleted by Twitter for violating its rules for healthy communication.

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

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 section. I first compile a list of violent keywords and filter the existing corpus through the keywords. The only difference is that the filtering on violent keywords is conducted on my own machine, not on the Twitter API. However, it can 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 representative set of keywords used to threaten one's partisan opponent (e.g., kill, shoot, choke, etc).

To deal with this, I combine model-based extraction of keywords with human judgment. 1) 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 threatening intention of violence or not. It is intended 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 from the Conversation AI team, a research initiative founded by Jigsaw, a unit within Google (Jigsaw 2020).[4] 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 intent in terms of the weights assigned to them. 2) Given the weighted terms, I then use human judgement 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 two rounds of filtering for violent Tweets is based on a list of violent keywords that people actually use in online space as well as consider to be violent and thus guarantee high level of coverage, only a small fraction of the violent-keyword Tweets actually contain intention of violence. This is because many Tweets are by no means violent but still contain a violent keyword in various ways. For instance, see how a violent keyword *shoot* is used in the following two Tweets: 1) Biden told people there how police should *shoot* unarmed people, liberals are way too easily impressed. 2) @realDonaldTrump Someone please *shoot* this bitch already. We can see that the keyword, shoot, is used to simply deliver

[4]The data set can be found *here*.

information about Biden's remark in 1) while it is used to actually promote violence in 2).

To handle this, I human-labeled a set of Tweets with a political and a violent keyword and built various machine learning classifiers. Specifically, I used active learning (Linder 2017, Miller, Linder, and Mebane 2020, Settles 2009). In active learning, we human-label a *randomly selected* texts, train a machine learning classifier, make predictions on unseen texts, *select (not randomly)* and human-label the texts whose label the classifier is most uncertain about (ones whose predicted probabilities are around the decision threshold), and finally accumulate the additionally human-labeled texts to re-train the classifier. Important is that the corpus of Tweets compiled through Step 1 and 2 is highly imbalanced data with only a small fraction containing a threat of political violence. 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 statistical learning model to learn about what features make a Tweet a threat of political violence (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

In this section, I provide a comprehensive set of data analyses with regard to the characteristics and spread of Tweets containing violent political rhetoric. The following data 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 the 2020 Presidential Election (campaigning, the election day itself, controversies over the election results) and culminates in the suspension of the former President Trump's Twitter account for his association with the Jan 6 Storming of the Capitol. As described in the previous section, I a) collected, through the Streaming API, Tweets have at least one political keyword, b) filtered though a set of violent keywords, c) then machine-classified them into Tweets that contain a threat of violence and ones that do not. The key findings include The key findings include a) Violent political rhetoric is rare (0.07% of political communication) but is spread through multiple chains on following ties (about 10% of retweets traveling over at least three ties), incidentally exposing a potentially huge size of audience and thereby amplifying its effects. b) users who write violent Tweets are on the fringe of the Twitter network and are ideologically more extreme and liberal than non-violent users. c) Spread of violent Tweets takes place primarily among ideologically-homogeneous users but there is also substantial amount of cross-ideological exposure.

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 and violent political Tweets, following a feature selection/weighting method for comparing word usage across different groups (Monroe, Colaresi, and Quinn 2008). In Figure 2, x-axis indicates the frequency of words across the type of Tweets (violent vs. non-violent). On the other hand, as we move up along the y-axis in each panel (non-violent on the top and violent at the bottom), words more frequently used by the type of Tweets are displayed. 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 language). Note that some of the words included as the words indicating violent political Tweets have already been baked in as part of the filtering by violent-keyword Tweets (see p. 8).

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 the violent terms involve many accounts that belong to high-profile political figures such as @realdonaldtrump (Donald Trump), @senatemajldr (Mitch McConnell), @mike_pence & @vp (Mike Pence), @secpompeo (Mike Pompeo), and the Squad (Alexandria Ocasio-Cortez, Ilhan Omar, Ayanna Pressley, and Rashida Tlaib), no entity-specific word is included in the top-3o list for non-violent Tweets. In particular, the account of the former President Trump, "@realDonaldTrump" is ranked second on the list, demonstrating that he and his online activities were at the center of violent and divisive communication on Twitter. The frequency of entity-specific words are also consistent with our focus on targeted violent political rhetoric. For the words indicating non-violent Tweets, we observe that general political terms are included (e.g., vote, ballot, state, tax, county, rally) along with words, such as "georgia", that represent a particular political event (the Senate race in Georgia and the main and runoff elections). [5]

[5]Two other characteristics of violent Tweets involve a) use of words that indicate intention/willingness (will, hope) and b) frequency of uncivil/insulting words (fuck, ass, shit).

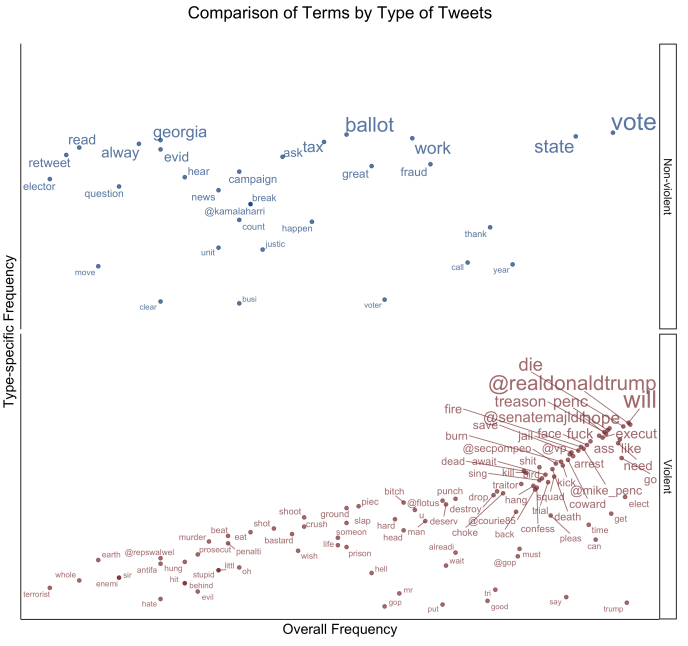


Figure 2. Comparison of Terms by Type of Tweets

Rank	Non-violent	Violent	Rank	Non-violent	Violent
1	vote	will	16	video	jail
2	ballot	@realdonaldtrump	17	sign	save
3	state	die	18	fraud	@vp
4	work	hope	19	great	arrest
5	georgia	penc	20	vaccin	go
6	tax	treason	21	michigan	coward
7	alway	@senatemajldr	22	campaign	burn
8	counti	fuck	23	economi	@secpompeo
9	read	execut	24	import	kick
10	evid	like	25	hear	shit
11	ralli	ass	26	elector	death
12	record	need	27	number	dead
13	retweet	face	28	fund	kill
14	ask	fire	29	approv	await
15	report	@mike_penc	30	question	squad

TABLE 1 *Comparison of Terms by Type of Tweets*

Now that we understand the stylistic characteristics of violent political rhetoric, what issues are talked about in Tweets containing such rhetoric? To provide a general sense of the content in such 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 I extracted hashtags to group hashtags that only differ in capitalization. In general, the hashtags together show that the content of violent political rhetoric is highly variegated, revolving around diverse 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 No.1 enemy of the country. Although it is beyond the scope of this study to look into the each of the hashtags and corresponding issues[6], they together make it clear that violent political rhetoric responds to and arises from various political/social issues.

Rank	Hashtag	Count	Rank	Hashtag	Count
1	#wethepeople	1511	16	#pardonsnowden	365
2	#1	1398	17	#traitortrump	358
3	#pencecard	1341	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

TABLE 2 *Most Frequent Hashtags in Violent Political Rhetoric (entire period)*

[6]Some will be discussed in the next section.

Timeline of Violent Political Tweets

Then, how frequent are Tweets containing violent political rhetoric? Figure 3 illustrates the timeline of Tweets with violent political rhetoric for our data collection period. The trend is expressed in terms of their count (absolute frequency) and of its 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 (i.e., intention of political violence, an extreme form 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 rhetoric in online political communication. For instance, Siegel, Nikitin, et al. (2019) reports that, during June 17, 2015 and June 15, 2017, any given month between 0.1% and 0.3% of Tweets contain contain hate speech. Although these violent Tweets comprises only a small fraction of political discussion, it is important to note that even these numbers amount to thousands of Tweets that threaten and incite violence, per day, and it is seen by the number of users engaged in political discussion that is far greater than that of the such Tweets themselves.

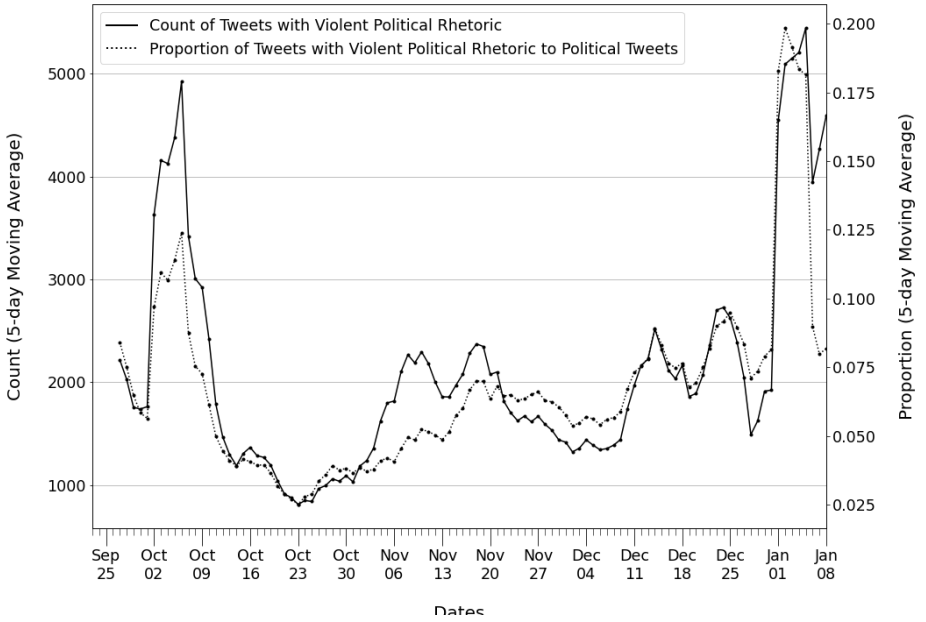


Figure 3. Timeline of Violent rhetoric in Political Tweets (23 Sep 2020 - 8 Jan 2021)

Then, what issues give rise to the outburst of violent political rhetoric on Twitter? 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 towards the election day (4 November 2020). To provide a general sense of what issues drive such trend, Table 3 reports the weekly top-5 hashtags included in Tweets of violent political rhetoric. While the steady uptrend towards 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. As for the earlier big spike, the hashtags for the week of 30 Sep through 6 Oct (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 10/2 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 the far-right and conspiracy theorists who attempt to delegitimize the election results in the wake of Trump's denial of his loss. 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.[7] It is also important to note that, while the general prevalence of violent political rhetoric in November and December reflects the partisan disputes over the election results (#treason, #diaperdon, #fightbacknow, #stopthestreal) along with other politically salient issues (e.g., death penalty in the week of 9-15 Dec), the drastic uptrend starting in the last week of 2020 appears to be predominantly driven by the partisan tension over the election disputes and the extremist discourse agitated by Trump's continuous mobilizing effort, inside and outside Twitter. Considering Trump's 26 December Tweet instigating his radical supporters to gather in D.C. on January 6th and the storming of the Capitol on that day[8], it is abundantly clear that offline political conflict is intertwined (potentially causally) with violent political rhetoric on Twitter.

[7]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).

[8]Trump Tweeted on 26 Dec 2020 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 6 Jan 2021, a joint session of Congress was scheduled to be held to count the Electoral College and to formalize Biden's victory.

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

TABLE 3 *Most Frequent Hashtags in Violent Political Rhetoric (weekly)*

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 (the same red line as in Figure 3) and of offline political violence (the new purple line).[9] Although this comparison based on crude aggregate measurement hardly provides 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

[9]The data for offline political violence is from U.S. Crisis Monitor (USCM) that is led by Armed Conflict Location and Event Data (ACLED) and the Bridging Divides Initiative (BDI) at Princeton University. USCM provides real-time data and on political violence in the U.S. For the count of violent political events, I include the following categories 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, go to *U.S. Crisis Monitor*.

the form of offline political violence (e.g., the storming of the Capitol on Jan 6 2021) and dangers it poses to democracy, it is crucial for furture research to investigate whether and how offline political violence translates into violent political rhetoric in the online political discussion (or vice versa).[10]

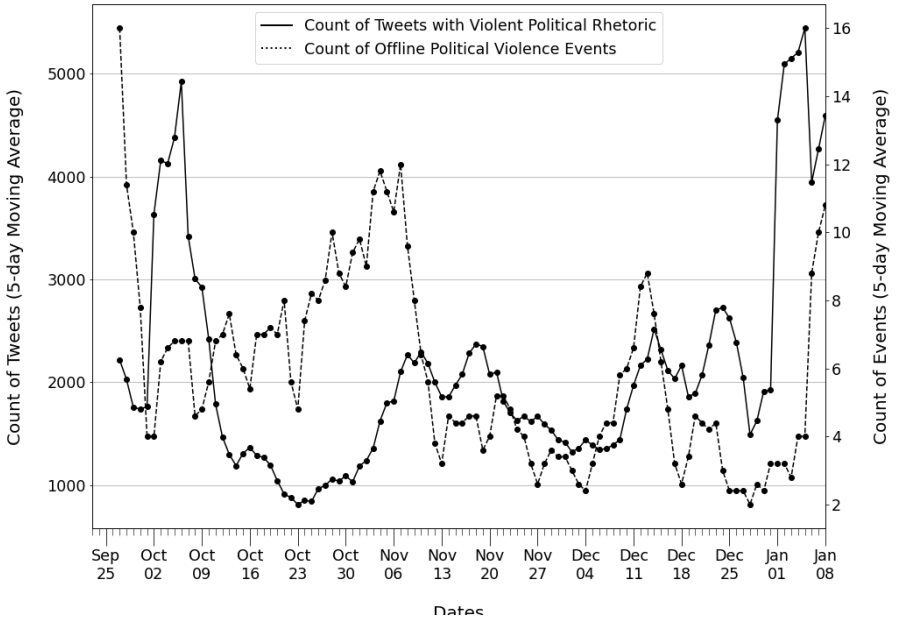


Figure 4. Timeline of Violent Political Rhetoric and Offline Political Violence (23 Sep 2020 - 8 Jan 2021)

[10]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).

Political Accounts in Violent Tweets

As previously discussed, in the era of radical/lethal partisanship, a substantial portion of citizens see political violence as acceptable and politicians in the opposition are often targeted with threats of violence in online space. What politicians appear in Tweets that contain threats of political violence? Are female politicians targeted more frequently than male ones? In terms of partisan affiliation, are Republican politicians targeted more often than their Democratic counterparts? **Tweets that contain violent political rhetoric in the data set "mention" a major politician's account, it often indicates the account in a given violent Tweet is the target of the threat.**

Politicians' account can be "mentioned" either by users directly including the account in a Tweet or by them replying to politicians' Tweets. Table 2, 3, and 4 report what political accounts appear in violent Tweets and present it by office, party, and gender. The two columns record, for different categories, the average number of violent Tweets that include political accounts in a given category. The first column is for the count of violent Tweets that are written as a reply to a Tweet written by a given political account while the second is for the count of violent Tweets that are written either as replies (as in the first column) or by independent violent Tweets.

First of all, Table 2 shows that Trump, the former president, is at the epicenter of violent partisan expressions on Twitter. As a single political figure, he appears in far more violent Tweets than all the other political accounts combined. The contender for presidency, Joe Biden, attracts the second largest number of violent Tweets followed by the incumbent vice president, Pence, and by vice president candidate from Democratic Party, Harris. Also, representatives, as opposed to governors and senators, are not the center of attention in violent political Tweets. Presumably, it might be due to the large number of representatives that make them less likely to get sufficient individualized attention to stimulate violent partisan expressions.

Given that the president, a Republican and male, can skew the comparison based on partisan affiliation and gender, Table 3 and 4 report the relevant statistics computed without violent Tweets that involve the president's account. Table 3 shows that Non-Republicans (including Democrats) appear more frequently than Democrats across all the statistics. These difference in terms of partisan affiliation is interesting considering the recent political climate of affective polarization and radical partisanship. In Table 4, we can see that, on average, male politicians appear more frequently in violent Tweets than female politicians. Given journalistic reports and academic study for online misogyny targeted at female politicians and ordinary female users (Vox Article, Felmlee, Rodis, and Zhang (2020), Fuchs and SchÄfer 2019), the difference in terms of gender is also worth more investigation.

To further explore how type of office, partisan affiliation, and gender are associated with the frequency of appearances of (thus the frequency of threats targeted against) politicians' accounts in violent Tweets, Table 5 shows the results from a Poisson regression analysis

where the count of mentions in violent Tweets, the response variable, is regressed against type of office, partisan affiliation, and gender. The results also include two covariates to account for the effect of politicians' accounts' popularity and activity in the Twitter network on the response variable: the number of followers and the number of Tweets published by a given political account.

The results reveal that, as opposed to what the descriptive statistics show, being Republican female is positively associated with appearance (and targetedness) in violent Tweets (the Full model). Why do Republican politicians appear and are targeted more often in violent Tweets than Democratic ones? One possibility is that politicians who belong to the party in power (i.e., Republicans) are more targeted

Although this might simply be a function of the proportion between liberal and conservative users on Twitter. As often pointed out in previous research, Twitter users are younger and more likely to be Democrats than the general population (Pew Research Center Report). This means that liberal users who outnumber conservative ones write more violent Tweets that target Republican politicians than their conservative counterparts do again Democratic politicians. However, further research is necessary as to why a particular group of politicians often become the targeted with violent political language on social media. In addition, in line with evidence for online misogyny, female politicians are more frequently targeted than their male colleagues and, as in the case of partisan differences, further investigation into why female politicians are more frequently targeted than their male colleagues.

	Mean Reply-to Count	Mean Mention Count
Trump	48,621	105,496
Pence	400	840
Biden	1,401	6,536
Harris	151	416
Governors	58	140
Senators	53	133
Representatives	14	33

TABLE 4 *Account Appearances by Office Type*

	Mean Reply-to Count	Mean Mention Count
Republicans	27	63
Non-Republicans	28	79

TABLE 5 *Account Appearances by Partisan Affiliation*

	Mean Reply-to Count	Mean Mention Count
Female	24	65
Male	29	74

TABLE 6 *Account Appearances by Gender*

	Position	Gender	Party	Follower	Tweets	Full
Pence	3.25***					−2.02***
	(0.04)					(0.04)
Biden	5.30***					−0.02
	(0.01)					(0.02)
Harris	2.55***					−1.66***
	(0.05)					(0.05)
Governors	1.45***					0.33***
	(0.01)					(0.02)
Senators	1.41***					−0.64***
	(0.01)					(0.01)
Female		−0.12***				0.22***
		(0.01)				(0.01)
Republican			−0.22***			1.19***
			(0.01)			(0.01)
Followers Count (log)				0.92***		1.04***
				(0.00)		(0.00)
Tweets Count (log)					1.32***	0.48***
					(0.01)	(0.01)
(Intercept)	3.48***	4.30***	4.37***	−6.91***	−7.31***	−13.15***
	(0.01)	(0.01)	(0.01)	(0.03)	(0.07)	(0.09)
AIC	137543.94	205192.50	204797.79	52456.90	164369.60	36943.89
BIC	137570.20	205201.25	204806.54	52465.58	164378.28	36987.28
Log Likelihood	−68765.97	−102594.25	−102396.89	−26226.45	−82182.80	−18461.95
Deviance	135637.02	203293.58	202898.86	50588.42	162501.12	35059.41
Num. obs.	588	588	588	566	566	566

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

TABLE 7 *Factors explaining Count of Violent Tweets Appearing in Political Accounts (Poisson Model)*

Engagement in Political Communication Network by Tweeter Type

How active are users who use violent language in the political communication network on Twitter? How so are they relative to ordinary users? This question is important because the more central to the network violent Tweeters are, the more likely ordinary users connected either directly or directly to violent ones are exposed to violent language. Table

6 and Figure 4 report the summary statistics and distribution for user-level indicators that measure engagement in the political communication network. Since some users exhibit disproportionately large values on these indicators, I report the median for Table 6 and the log-transformed distribution for Figure 4. 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 and spread of their content can be generally limited than non-violent political users'. In sum, violent users tend to be located in the fringe of the political communication network on Twitter.

Count	Violent	Non-violent
Friends	195	446
Followers	52	206
Likes	2,236	6,799
Tweets	1,795	6,028

TABLE 8 *Median Value for Major User Characteristics*

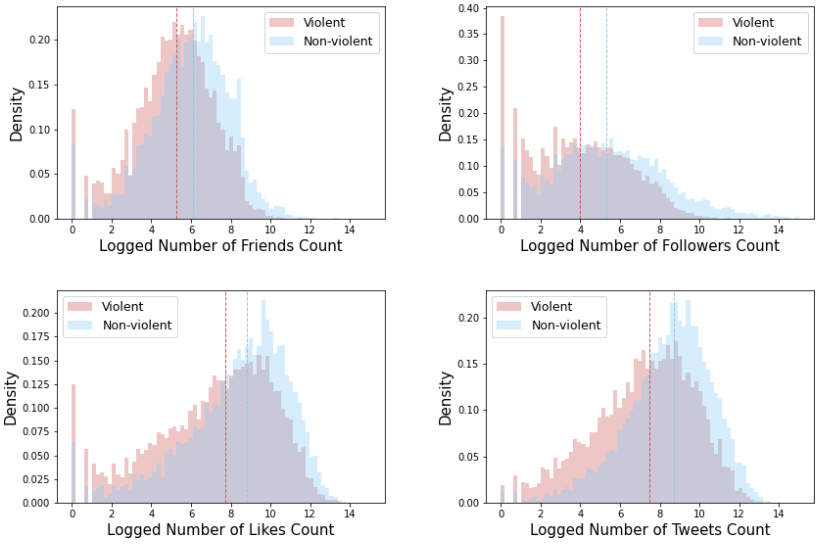


Figure 5. Distribution for Network Engagement Indicators

Distribution of Ideology by Tweeter Type

Figure 5 shows the distribution of ideology scores for violent and non-violent Tweeters. I estimated users' ideology using an ideal point estimation approach introduced in Barberá (2015). The score is computed to center around 0 (the ideological center) with negative numbers indicating liberal and positive numbers indicative conservative.

First, the distribution of non-violent Tweeters shows that they tend to be more liberal. This is consistent with the conventional wisdom that Twitter users tend to be liberal, younger, and Democrats (Pew Research Center Report).[11] Second, what is noteworthy is that violent Tweeters are slightly more liberal than non-violent Tweeters. We can see that the mean ideology score of violent Tweeters leans more toward the liberal direction than that of conservative Tweeters. The result of Welch two-sample t-test shows that the difference is 0.16 and is statistically significant (95% C.I.: 0.13, 0.19). In other words, in the data set, liberal Tweeters are more violent than conservative Tweeters.

However, the difference is quite small and there is over-time heterogeneity. Figure 6 shows that, while violent users tend to be more liberal than non-violent ones for the first seven weeks, the trend is flipped for the last five weeks. Given the Presidential Election and counting of the votes took place during the week 7 (Nov 4 - Nov 10, 2020), this is likely that the election marks a turn point for the ideological trend of violent political language on Twitter. These findings also tell us that use of violent language in online space is unlikely to be stably related to ideology but reflects how particular phrases of politics (including elections, protests, etc) stimulate violent partisan expressions.

Next, Figure 7 shows that violent Tweeters are more ideologically extreme than non-violent Tweeters. To calculate the ideological extremity score, I took the absolute value of the ideological scores used for in Figure 5 (ranging from -4 to +4) to get a sense of how one is far apart from the middle point of the ideological continuum (indicated with 0). Importantly, the same results are found in almost all of the weekly distributions in Figure 8.

[11] Since the vast majority of political Tweeters are non-violent ones, the distribution for non-violent Tweeters is nearly identical to the distribution for political Tweeters.

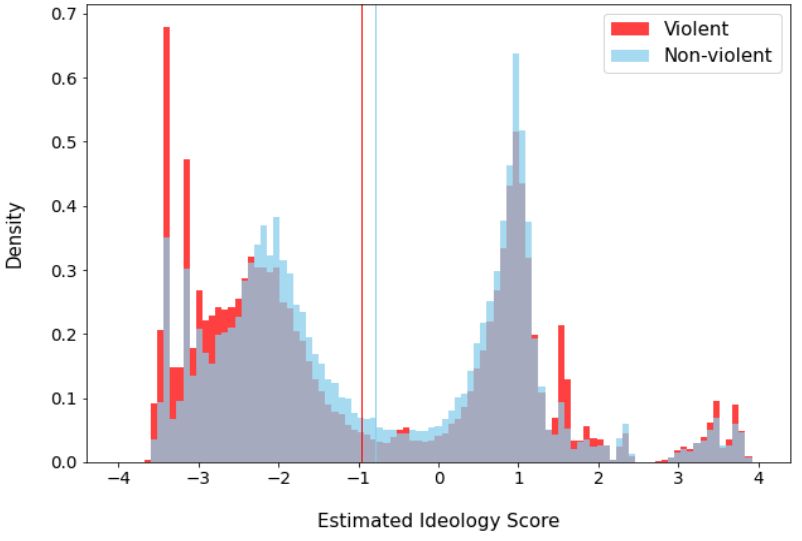


Figure 6. Distribution of Ideology by Type of Political Tweeters

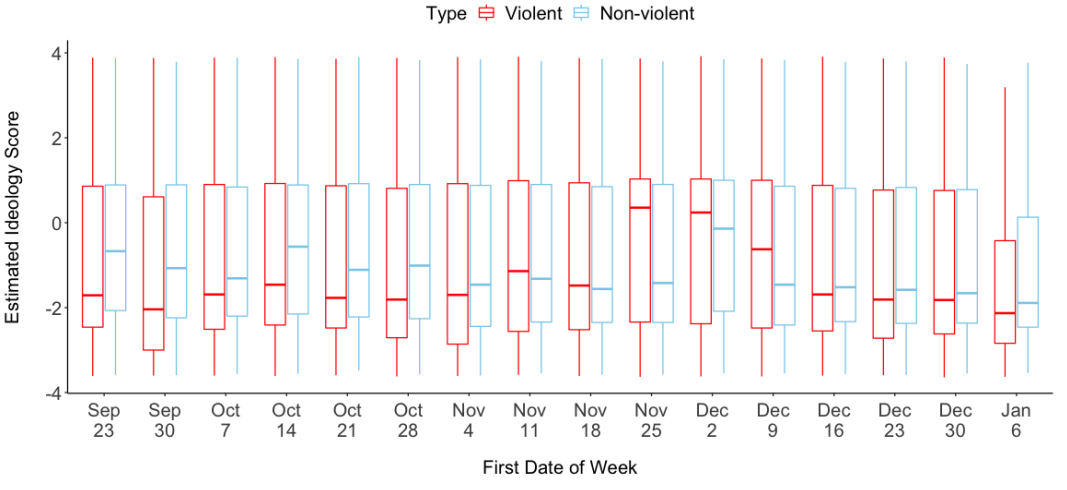


Figure 7. Weekly Distribution of Ideology by Type of Political Tweeters

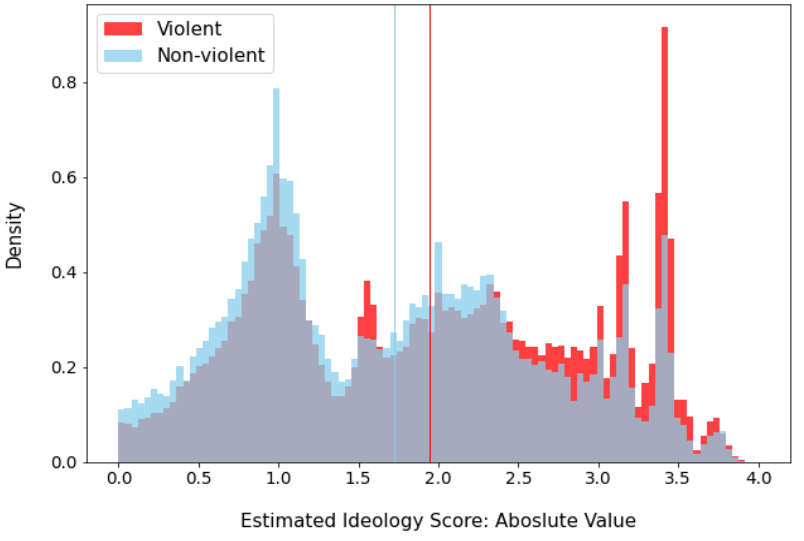


Figure 8. Distribution of Ideology by Type of Political Tweeters

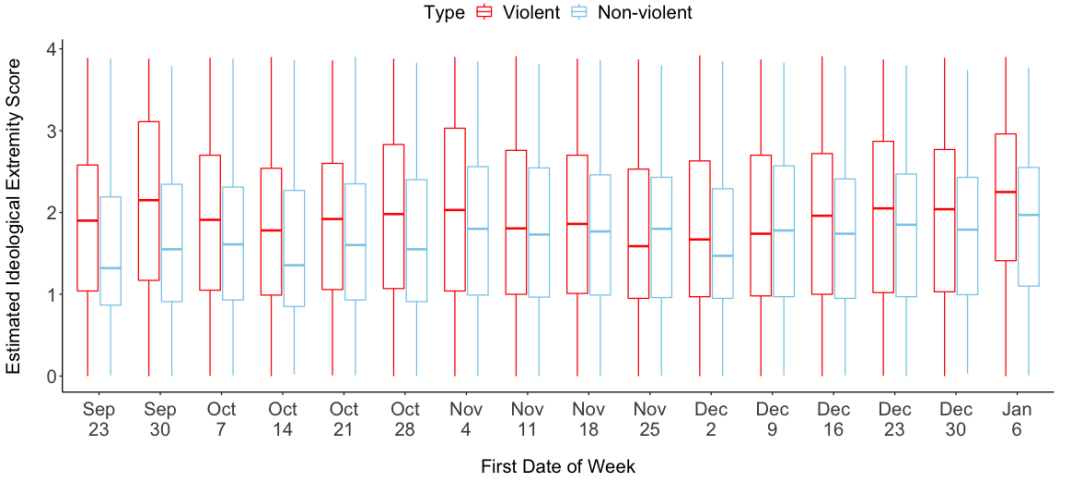


Figure 9. Weekly Distribution of Ideology by Type of Political Tweeters

Spread of Violent Political Rhetoric

How do Tweets containing violent political rhetoric spread and how far? First, existing research on political communication on social media suggests that, while political information is exchanged primarily among individuals who are ideologically homogeneous (Barberá et al. 2015), there is also 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 9 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 respective Pearson's R scores are 0.69 (violent) and 0.72 (non-violent). Thus, if we compare the level of ideological homophily — the tendency for individuals to form ties with those who are ideologically close to themselves — in retweeting to what would have happened in its absence, we can see a) that in general political retweeting is strongly affected by ideological homophily and b) ideologically homophily also applies to violent political rhetoric.

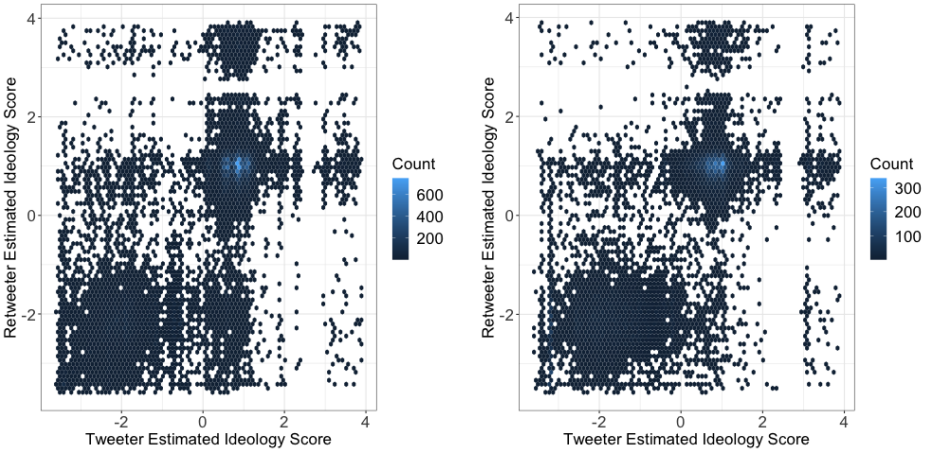
These findings implies that, compared to non-violent political Tweets, violent ones are no more confined to the "ideological echo-chamber." Although the results shows that the spread of violent political rhetoric on Twitter takes place primarily among like-minded users but there is as much cross-ideological exposure as there is for non-violent (ordinary) political communication. Given the considerable amount of cross-ideological exposure to violent political rhetoric, it is important for future research in political communication to investigate how (differently) exposure to in-party violent rhetoric and exposure out-party violent rhetoric affect various political behavior/attitudes.

Next, how far do violent Tweets travel on the Twitter communication network? As in the previous Table 6 and Figure 4, violent Tweeters tend to lie on the fringe of the Twitter political communication network. However, their content still can travel to a large size of audience through indirect ties. Figure 10 describes the distribution of the shortest path distance on the following network for all the retweets in the data set. The distance was estimated as 1 if the Retweeter is in the the Tweeter's followers list or the Tweeter is in the Retweeter's friends list. In a similar manner, the distance was 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 for distance 1 nor the condition for distance 2 is met, the shortest distance is coded as 3 or more.

As seen in the figure, almost two thirds of the retweets take place between users with a direct tie (64%). However, there is a substantial minority of Tweets that travel over the Tweeter's follower ties. Around 28% of the retweets take place between users whose estimated shortest path distance is 2. For the remaining 9%, Tweets were retweeted at least over three ties.

Given the findings that even exposure to mild violent rhetoric has crucial political

consequences (Kalmoe 2014; Kalmoe, Gubler, and Wood 2018; Kalmoe 2019), these findings are impressive in terms of the reach of (or magnitude of the exposure to) violent Tweets. That is, even if one does not follow a violent Tweeter, it is possible for one to get exposed to such discomforting political content against one's intent. In addition, the impact of violent Tweets can be 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 and creating another opportunity for the violent Tweet to be retweeted.



(a) Violent Tweets

(b) Non-violent Tweets

Figure 10. Ideological Homophily: Correlation between Tweeters' and Retweeters' Ideology

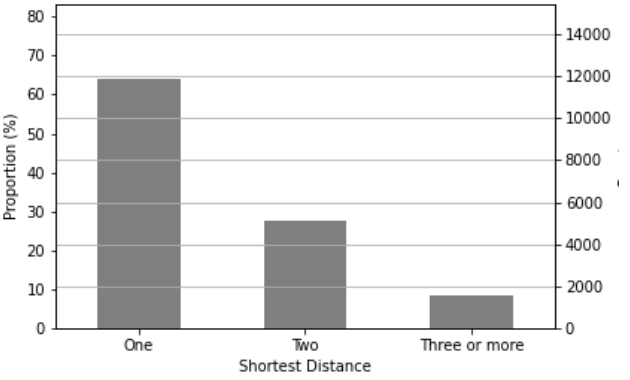


Figure 11. Comparison of Terms by Type of Political Tweets

APPENDIX

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) annotated Tweets in terms of whether a given Tweet contains a threat of political violence or not. Note that Tweets presented for labeling are (already assumed to be) political and the coders were tasked with identifying a threat of violence in a given Tweet. The Tweets were presented as reformatted on Google Sheet. A Tweet that is a quote of another Tweet is presented with the quoted Tweet because the meaning of the former is much 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 an 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 involves 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 a threat of real violence (e.g., 'burn in hell!'), 3) mentioning (or even criticizing) a threat of violence uttered by someone else, and 4) sarcasm (e.g., 'why don't just shoot them all if you believe violence solves the problem?'). Detailed coding guidelines for dealing with such confusing cases are documented in *this Github repository*.

The coders manually annotated a set of 2,500 Tweets together (meaning each Tweet is annotated three times in total). Specifically, the coders worked on the initial 2,000 Tweets together to refine coding guidelines and then manually annotated another 500 Tweets. Again, after manually annotating the 500 Tweets, the coding guidelines were updated. Then, with the coding guidelines built based on the 2,500 Tweets were used for later manual annotation of another set 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 annotated.

As previously noted, the concept of threats of violence is inherently vague and subjective. Accordingly, the level of inter-coder agreement for studies on similar concepts is generally moderate (Table 1). The inter-coder agreement scores achieved in our study are reported in Table 2. 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 relevant literature.

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

TABLE A1 *Inter-coder Agreement in Research on Similar Concepts*

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	

TABLE A2 *Inter-coder Agreement from the 500 Manually-annotated Tweets*

Appendix B. Active Learning and Classification

Based on the literature on active learning (see Linder 2017), 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 data set of Tweets containing political and violent keywords (C_{pv}).

2. Including myself, three human annotators label the selected Tweets in terms of whether a given Tweet contains a threat of violence or not. A machine learning classifier is trained on the annotated Tweets.

3. Next, the trained classifier is fit on the rest of C_{pv1} and the predicted probability of being violent is calculated for each of the Tweets.

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

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

For the first round, I randomly sampled 2,500 Tweets and annotated 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 annotated 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.[12] Through this iterative process, a total of 10,000 Tweets containing political and violent keywords are manually annotated.

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 for evaluating the performance of the classifiers. The results are reported in Table 3. 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 far upper part of the probability distribution

Note that the model's performance parallels the performance achieved in 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 the same 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 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).

Model	Precision	Recall	F-1
Logistic Regression + Count Vector	69.25	35.08	46.52
Logistic Regression + TF-IDF Vector	81.80	10.12	18.01
Logistic Regression + GloVe	60.85	11.13	18.82
Random Forest + Count	77.66	19.99	31.76
Random Forest + TF-IDF Vector	82.97	17.90	29.43
Random Forest + GloVe	74.89	11.38	19.69
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

TABLE A3 *The Average Performance of Machine Learning Classifiers from 5-fold Cross Validation*

[12]For detailed information about BERT, see **devlin2018bert**

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 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/.

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.

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.*

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.

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.

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.*

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/.

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.

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.

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.

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."

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.

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).*