Appendix

A. Selection of Political Keywords

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

While "mention" is a central feature around communication on Twitter, another way to engage politicians is simply using names (e.g., "Donald Trump" as opposed to "@realDonaldTrump"). While I considered using both accounts and names in filtering live tweets, it was impossible due to the limit to the number of keywords for filtering (Twitter 2021c). In addition, using names as opposed to accounts is prone to measurement error for many reasons. First, users call politicians by different versions of their names, making it difficult, if not impossible, to decide on a particular version for each politician. For instance, there are cases where politicians are called by the last name only, the full name, or various abbreviations (e.g., TJ Cox). Also, there are issues related to homonyms for many politicians (e.g., the North Carolina Representative David Price and the baseball pitcher David Price).

Nevertheless, it is crucial to examine whether data generated by the list of politicians' accounts differ from data generated by their names. This is because systematic differences

between the two in terms of key dimensions of the downstream analysis (e.g., political party, gender) might introduce bias. To do so, I started with collecting tweets including any of the politicians' full names. Then, I calculated the proportion of the number of tweets including a given politician's full name to the number of tweets including the politician's account. Finally, I compared the median proportion across major politician-level attributes highlighted in my substantive analysis.

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

Figure A1 depicts the distribution of the proportion of the number of full name tweets to the number of mention tweets (expressed in percentage). The original distribution is highly skewed so I log-transformed it. The figure shows that most of the observations are concentrated in the area left to the 100% point at which the numbers of full name tweets and of mention tweets are equal. Because the distribution is skewed, I used the median for the central tendency measure. The median proportion, 13.04%, indicates that only a small fraction of tweets engaging politicians on Twitter use their full names as opposed to their accounts. In addition, Table A1 breaks down the median proportion across gender, political party, and position. We can see that there are no noticeable discrepancies in the proportion across the three characteristics. This provides evidence that tweets including

politicians' full names and tweets including their accounts are not systematically different with regard to the key dimensions of comparison in the substantive analysis.

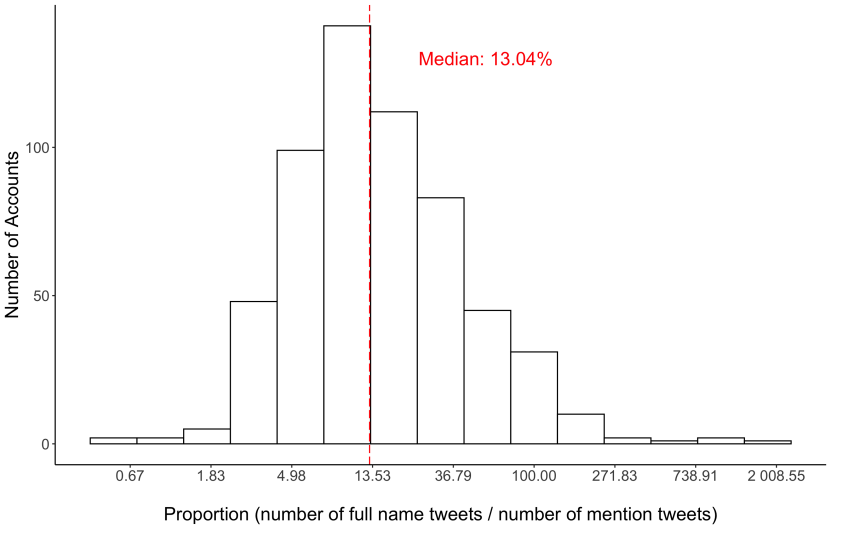

Figure A1. Distribution of the proportion of full-name tweets to mention tweets)

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

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

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

B. Selection of Violent Keywords

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

It is important to note that the set of keywords I extract from the corpus is highly

comprehensive. I extract 200 uni- and bi-gram keywords that are the most predictive of perceived threat. I experimented with different threshold values (e.g., 100, 200, 300, 400) and set the threshold based on my judgment of the point beyond which keywords are no longer meaningfully associated with perceived threat and are likely to only result in false positives. Since I manually label the resulting violent-keyword tweets (tweets filtered in through the 200 violent keywords) in terms of whether the tweet is actually violent in the next step (Step 3), I was able to include even keywords that are marginally associated with threat perception. The list not only involves a wide variety of violent keywords (e.g., die, punch, choke) but also cover their semantic variants (e.g., "die", "dead", "death"). Furthermore, the list involves many keywords that are not necessarily violent themselves but are often used in violent rhetoric (swear words, auxiliary verbs, or collocative structures).

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

C. Manual Labeling

Three human coders (including myself and two undergraduate assistants) labeled tweets in terms of whether a given tweet contains violent political rhetoric or not. Specifically, the human coders were instructed to classify tweets into three classes. Class 1 is “violent politic rhetoric" where the author expresses the intention of physical harm against a political opponent. Class 2 is “violent political metaphor" where the author’s statement about essentially non-violent politics is expressed using a violent metaphor but still lacks a violent intention. While Class 2 is not directly related to my study, tweets that fall into this class appear frequently enough to constitute a separate class. Class 3 is a garbage can class for tweets that are neither Class 1 nor Class 2. Tweets were presented on Google Sheet. A tweet that quotes another tweet is presented with the quoted tweet because the former’s meaning is more clear with the latter.

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

The coders manually labeled a set of 2,500 tweets together (meaning each tweet is labeled three times). Specifically, the coders worked together on the initial 2,000 tweets to refine coding rules and manually labeled another 500 tweets. After the 500 tweets, the coding rules were updated again. Then, the coding rules based on the 2,500 tweets

were used for later manual annotation of 7,597 tweets. For the 7,597 tweets, three coders worked on three different sets of tweets (Coder 1: 3,500, Coder 2: 3,500, Coder 3: 597). In sum, a total of 10,097 tweets were manually labeled.

As previously noted, the concept of violent political rhetoric (and aggressive speech in general) is inherently subjective. Accordingly, the levels of inter-coder agreement reported in studies on aggressive speech are low to moderate (Table A2). In Table A3, I report the inter-coder agreement scores in my study. It shows that, by any measure, the level of inter-coder agreement outperforms the standard in the relevant literature.

TABLE A2 *Inter-coder agreement on similar concepts*

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

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

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

D. Active Learning and Machine Classification

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

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

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

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

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

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

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

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

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

While the BERT model performs well, note that the model inevitably makes errors. It is particularly the case for tweets discussing the use of violence. For instance, discussion of the death penalty (e.g., the case of Brandon Bernard) tends to involve many violent

expressions (e.g., kill, death, die, etc.) and classifying tweets in this context can be a challenging task for any machine learning model. The BERT model still successfully identifies violent political rhetoric arising from such discussion (e.g., *"@realDonaldTrump If you don't stop the execution of Brandon Bernard I hope you die a very painful death. It is the least you deserve you POS!"* or *"@realDonaldTrump @Varneyco I hope you catch an illness and die you orange turd it should've been you and Kyle Rittenhouse that should be injected with poison not Brandon Bernard I hope the White House burns down with you in it"*). At the same time, however, it can and do misclassify tweets simply discussing (or opposing) the prisoner being executed as violent (e.g., *"@realDonaldTrump BASTARD WHYD U N UR TEAM KILL EXECUTE BRANDON BERNARD"* or *"Brandon Bernard will be executed on HumanRightsDay"*).

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

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

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

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

E. Top-30 Keywords by Type-specificity

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

TABLE A6 *Comparison of terms by type of tweets*

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

F. Regression Analysis on Mentioning

Table A6 reports descriptive statistics for the mentioning analysis. Tables A7 and A8 report two additional models to assess whether the findings in the main text are robust to model specifications. The first model is the same as the main model but includes three candidates for the Presidential Election: Biden, Pence, Harris (except for Trump who is overly influential). The second model is a zero-inflated negative binomial model to account for excess zeros (the first-stage model uses the same set of variables as the second-stage model). Negative binomial regression is used for all of the models to deal

with over-dispersion. As seen in the coefficients, the results for position, gender, and partisan affiliation are consistent across the models.

TABLE A7 *Descriptive statistics for mentioning analysis*

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

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

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

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

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

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

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

G. Network Engagement Indicators

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

TABLE A10 *Median value for network engagement indicators*

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

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

Figure A2 depicts the distribution of ideology by type of tweeters (violent vs. non-violent), without tweets that mention Trump's account. Although the difference decreases to 0.09 on the ideological 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).

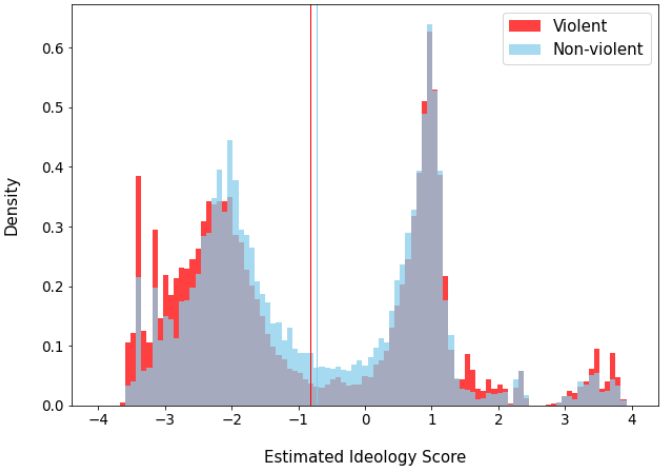

Figure A2. Distribution of ideology by type of political tweeters (without Tweets mentioning '@realDonaldTrump')

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