Journal of Computational Social Science (2022) 5:1–18
https://doi.org/10.1007/s42001-021-00117-9

RESEARCH ARTICLE

#Election2020: the first public Twitter dataset on the 2020 US Presidential election

Emily Chen[1] · **Ashok Deb**[1] · **Emilio Ferrara**[1]

Received: 7 March 2021 / Accepted: 19 March 2021 / Published online: 2 April 2021

Abstract

Credible evidence-based political discourse is a critical pillar of democracy and is at the core of guaranteeing free and fair elections. The study of online chatter is paramount, especially in the wake of important voting events like the recent November 3, 2020 U.S. Presidential election and the inauguration on January 21, 2021. Limited access to social media data is often the primary obstacle that limits our abilities to study and understand online political discourse. To mitigate this impediment and empower the Computational Social Science research community, we are publicly releasing a massive-scale, longitudinal dataset of U.S. politics- and election-related tweets. This multilingual dataset encompasses over 1.2 billion tweets and tracks all salient U.S. political trends, actors, and events from 2019 to the time of this writing. It predates and spans the entire period of the Republican and Democratic primaries, with real-time tracking of all presidential contenders on both sides of the aisle. The dataset also focuses on presidential and vice-presidential candidates, the presidential elections and the transition from the Trump administration to the Biden administration. Our dataset release is curated, documented, and will continue to track relevant events. We hope that the academic community, computational journalists, and research practitioners alike will all take advantage of our dataset to study relevant scientific and social issues, including problems like misinformation, information manipulation, conspiracies, and the distortion of online political discourse that has been prevalent in the context of recent election events in the United States. Our dataset is available at: https://github.com/echen102/us-pres-elections-2020.

Keywords Presidential election · Twitter · Social media analysis

✉ Emily Chen
echen920@usc.edu

Ashok Deb
adeb@usc.edu

Emilio Ferrara
emiliofe@usc.edu

[1] Information Sciences Institute, University of Southern California, 4676 Admiralty Way, #1001, Marina del Rey, CA 90292, USA

Introduction

In 2020, Americans returned to cast their vote for the next president of the US: incumbent Republican Donald J. Trump or the Democratic challenger, and former Vice-President Joseph R. Biden. We began collecting tweets in May 2019 in an effort to capture online chatter surrounding this defining democratic process and to make this collection available to the research community.

Historically, the incumbent president is favored to win their party's nomination for president;[1] although Trump did face a few challengers from the Republican party, it became increasingly clear that he would gain the Republican party's nomination.

Joe Biden officially accepted the Democratic nomination during the Democratic National Convention.[2] Donald Trump officially accepted his nomination on August 27, 2020, during the Republican National Convention.[3]

As the final sprint to election day on November 3, 2020 began, Americans took to online social platforms to voice their opinions and engage in conversation surrounding the elections. Twitter has historically been a platform used by politicians to reach their base [10], and has recently begun more aggressive efforts to tag posts as misleading and potentially incorrect in order to mitigate the spread of misinformation that had already been prevalent on the platform [4].[4] On election day, many again used social media to express their thoughts on the unfolding elections. News outlets were unable to call the elections for several days after election day, as many key states were still counting ballots; social media was used as a means to spread information (both factual and misleading) and to both protest and advocate for controversies surrounding ballots and the influx of mail-in ballots caused by COVID-19.[5,6]

On November 7, the media was finally able to call the election and named Biden as the president-elect, and Kamala Harris as the vice-president-elect.[7] Yet, in the aftermath of this pronouncement and in the current polarized nature of the United States political landscape, social media has become an environment where misinformation and disinformation can flourish and spread. President Trump refused to concede the election, and continued to promote the claim that the election had been stolen.[8,9] These claims from Trump bolstered the basis for the "stop the steal"

[1] https://time.com/5682760/incumbent-presidents-primary-challenges/.
[2] https://www.nbcnews.com/politics/2020-election/biden-set-accept-democratic-nomination-final-night-convention-gets-underway-n1237559.
[3] https://abcnews.go.com/Politics/rnc-2020-day-trump-accept-nomination-white-house/story?id=72577769.
[4] https://www.npr.org/2020/10/09/922028482/twitter-expands-warning-labels-to-slow-spread-of-election-misinformation.
[5] https://projects.fivethirtyeight.com/election-results-timing/.
[6] https://www.nytimes.com/2020/11/06/business/media/election-call.html.
[7] https://www.poynter.org/fact-checking/2020/what-does-it-mean-when-the-media-calls-an-election/.
[8] https://www.cbsnews.com/news/trump-tweet-claims-he-won-election-twitter-flags/.
[9] https://www.politifact.com/factchecks/2020/dec/14/facebook-posts/joe-biden-won-presidential-election-legally/.

campaign, and eventually culminated in a riot at the United States Capitol on January 6, 2021.[10,11] This led Twitter and other social media platforms to either semi-permanently or permanently suspend President Trump's accounts from their services, citing the riot and the potential for further incitement of violence as grounds for the bans.[12] Many vendors began to cut ties with right-wing social media platform Parler due to the role it played in coordinating the January 6 riot.[13] President Biden was inaugurated into office on January 20, 2021, along with Vice President Harris.[14]

Inspired by the positive impact that our similar initiative to share a COVID-19 Twitter dataset has had on the research community [3], in this paper, we document the release of our 2020 US Presidential election-related dataset that we have been collecting for over one year, a period covering all the events described above and more. We hope that, in releasing this dataset, the research community can leverage its content to study and understand the dynamics in a highly contentious election held during a pandemic. This dataset enables researchers to directly study the impact that the pandemic has had not only on the political landscape, but also on misinformation, disinformation and coordinated actors, with reports of confirmed foreign interference attempts already surfacing [7].[15]

Data collection

Data collection method

We uninterruptedly collected election-related tweets beginning *May 20, 2019*, and have continued collection efforts since then. We use Twitter's streaming API through the Tweepy library and follow specific mentions and accounts related to candidates who were running to be nominated as their party's nominee for president of the United States, in addition to a manually-compiled, general election-related list of keywords and hashtags.[16] As candidates officially announced the suspension of their campaigns, their respective accounts and mentions were removed from our real-time tracking list. In response to real-world events, we decided to restart tracking for a subset of these accounts, in addition to adding supplemental keywords and accounts to our tracking list. This is documented in Table 1.

We will continue to collect election-related tweets at least through the first six months of the Biden administration, so as to capture the nation's post-election and

[10] https://www.npr.org/sections/live-updates-2020-election-results/2020/11/08/932543826/the-next-2020-election-fight-convincing-trumps-supporters-that-he-lost.

[11] https://www.politifact.com/article/2021/jan/11/timeline-what-trump-said-jan-6-capitol-riot/.

[12] https://www.axios.com/platforms-social-media-ban-restrict-trump-d9e44f3c-8366-4ba9-a8a1-7f3114f920f1.html.

[13] https://www.bloomberg.com/news/articles/2021-01-10/apple-removes-parler-from-app-store-after-use-in-capital-riot.

[14] https://www.nytimes.com/2021/01/20/us/politics/biden-president.html.

[15] https://home.treasury.gov/news/press-releases/sm1118.

[16] https://www.tweepy.org/.

post-transition activity. In total, our dataset comprises well over 1 billion tweets. Release v1.12 contains 1,258,209,617 tweets, spanning from 12/01/2020 through 1/22/2021. In our latest (v1.16) and future releases, we will continue processing and adding data we collected prior to 12/01/2020 and after 1/22/2021.

Note: Twitter's Developer Agreement & Policy stipulates that we are unable to share any data specific to individual tweets except for a tweet's Tweet ID. As a result, we are releasing a collection of Tweet IDs that researchers are then able to use in tandem with Twitter's API to retrieve the full tweet payload. We recommend using tools such as DocNow's Hydrator[17] or Twarc[18]; if tweets have been deleted from Twitter's platform, researchers will be unable to retrieve the payloads for those tweets. We provide ready-to-use Python code scripts to perform all the operations described above in our repository.

Tracked keywords and accounts

In order to capture the chatter surrounding the 2020 US presidential elections, we followed specific user mentions and accounts that were and are tied to the official and personal accounts of candidates who ran for president. Twitter's streaming API gives us access to approximately 1% stream of all tweets in real-time, and takes in a list of keywords, returning any tweet within that sample stream that contains any of the keywords in the metadata and text of the tweet payload.[19] Thus it is unnecessary to track every permutation of each keyword. We list a sample of the mentions and accounts that we tracked in release v1.12 in Table 1 and a sample of the keywords we tracked in Table 2. A full list can be found in the accounts.txt file and keywords.txt file in our data repository.

Data and access modalities

We upgraded our data collection pipeline on June 20, 2020 for data collection reliability purposes. Data prior to June 20, 2020 experienced higher rates of technical collection issues. While our most recent release is Release v1.16, containing *1,355,356,627* tweets from December 1, 2019 through February 19, 2021, we focus on and detail release v1.12 throughout this study.

Release v1.12 (January 25, 2021)

Release v1.12 includes tweets collected from December 1, 2019 through January 22, 2021, containing *1,258,209,617* tweets in all. We are still continuing our computational efforts to pre-process and clean the rest of our existing dataset, and will be

[17] https://github.com/DocNow/hydrator.
[18] https://github.com/DocNow/twarc.
[19] https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/introduction.

Table 1 A sample of the mentions and accounts that we actively tracked (v1.12 — January 25, 2021)

Mentions	Started tracking	Stopped	Restarted
@realDonaldTrump	5/20/19	–	–
@GovBillWeld	5/20/19	–	–
@MarkSanford	5/20/19	11/14/19	9/25/20
@WalshFreedom	5/20/19	–	–
@MichaelBennet	5/20/19	–	–
@JoeBiden	5/20/19	–	–
@CoryBooker	5/20/19	1/13/20	9/25/20
@GovernorBullock	5/20/19	12/2/19	9/25/20
@PeteButtigieg	5/20/19	–	–
@JulianCastro	5/20/19	1/2/20	9/25/20
@BilldeBlasio	5/20/19	11/14/19	9/25/20
@JohnDelaney	5/20/19	–	–
@TulsiGabbard	5/20/19	–	–
@gillbrandny	5/20/19	11/14/19	6/20/20
@KamalaHarris	5/20/19	12/3/19	6/20/20
@SenKamalaHarris	5/20/19	12/3/19	6/20/20
@Hickenlooper	5/20/19	11/14/19	9/25/20
@JayInslee	5/20/19	11/14/19	9/25/20
@amyklobuchar	5/20/19	–	–
@SenAmyKlobuchar	5/20/19	3/3/20	6/20/20
@WayneMessam	5/20/19	12/2/19	9/25/20
@sethmoulton	5/20/19	11/14/19	9/25/20
@BetoORourke	5/20/19	11/14/19	9/25/20
@TimRyan	5/20/19	11/14/19	9/25/20
@BernieSanders	5/20/19	–	–
@ericswalwell	5/20/19	11/14/19	9/25/20
@ewarren	5/20/19	–	–
@SenWarren	6/20/20	–	–
@marwilliamson	5/20/19	–	–
@AndrewYang	5/20/19	–	–
@JoeSestak	5/20/19	12/2/19	9/25/20
@MikeGravel	5/20/19	8/6/19	9/25/20
@TomSteyer	5/20/19	–	–
@DevalPatrick	5/20/19	–	–
@MikeBloomberg	5/20/19	–	–
@staceyabrams	6/20/20	–	–
@SenDuckworth	6/20/20	–	–
@TammyforIL	6/20/20	–	–
@KeishaBottoms	6/20/20	–	–
@RepValDemings	6/20/20	–	–
@val_demings	6/20/20	–	–
@AmbassadorRice	6/20/20	–	–
@GovMLG	6/20/20	–	–

Table 1 (continued)

Mentions	Started tracking	Stopped	Restarted
@Michelle4NM	6/20/20	–	–
@SenatorBaldwin	6/20/20	–	–
@tammybaldwin	6/20/20	–	–
@KarenBassTweets	6/20/20	–	–
@RepKarenBass	6/20/20	–	–
@Maggie_Hassan	6/20/20	–	–
@SenatorHassan	6/20/20	–	–
@GovRaimondo	6/20/20	–	–
@GinaRaimondo	6/20/20	–	–
@GovWhitmer	6/20/20	–	–
@gretchenwhitmer	6/20/20	–	–

uploading batches of past and future data as they become available. A sample of the mentions/accounts and keywords that we followed can be found in Tables 1 and 2, respectively, with full lists of both available on our Github repository. Furthermore, Table 3 shows the top 40 most popular hashtags, grouped by general categories. We can clearly see that most of the hashtags are directly related to party campaigns and conspiracy theories surrounding the elections. Others are related to political events, social movements and the COVID19 pandemic.

As this dataset was curated for the 2020 US Presidential election cycle, it is unsurprising that the majority of these tweets are in English (see Table 4 for a breakdown of the languages in release v1.12).

Data access

The dataset is publicly available and continuously maintained on Github at this address: https://github.com/echen102/us-pres-elections-2020.

The dataset is released in compliance with the Twitter's Terms & Conditions and the Developer's Agreement and Policies.[20] This dataset is still presently being collected and will be periodically updated on our Github repository. Researchers who wish to use this dataset must agree to abide by the stipulations stated in the associated license and conform to Twitter's policies and regulations.

Data analysis

Although we are continuing to collect tweets to add to our data collection as we follow the transition to the Biden-Harris administration, we first present an analysis on tweets from our dataset from January 2020 through the end of December 2020. This

[20] https://developer.twitter.com/en/developer-terms/agreement-and-policy.

Table 2 A sample of keywords that we actively tracked in our Twitter collection (v1.12 — January 25, 2021)

Keywords	Tracked since
ballot	6/20/20
mailin	6/20/20
mail-in	6/20/20
mail in	6/20/20
donaldtrump	9/12/20
donaldjtrump	9/12/20
donald j trump	9/12/20
donald trump	9/12/20
don trump	9/12/20
joe biden	9/12/20
joebiden	9/12/20
biden	9/12/20
mike pence	9/12/20
michael pence	9/12/20
mikepence	9/12/20
michaelpence	9/12/20
kamala harris	9/12/20
kamala	9/12/20
kamalaharris	9/12/20
trump	9/13/20
PresidentTrump	9/13/20
MAGA	9/13/20
trump2020	9/13/20
Sleepy Joe	9/13/20
Sleepyjoe	9/13/20
HidenBiden	9/13/20
CreepyJoeBiden	9/13/20
NeverBiden	9/13/20
BidenUkraineScandal	9/13/20
DumpTrump	9/13/20
NeverTrump	9/13/20
VoteRed	9/13/20
VoteBlue	9/13/20
RussiaHoax	9/13/20

enables us to examine political discourse on Twitter through the Presidential primaries, debates and election. Highly political divisions have emerged in COVID-19 discourse [9], alongside conspiracy theories [6] and public heath related trends that have emerged due to COVID-19 [3]. Our recent work on this dataset has also shown that partisan trends drive the discourse on Twitter, with conservative users posting at much higher volumes compared to their liberal counterparts. Conservative users also tended to share more known conspiracy-related narratives [7]. We have also

Table 3 Top 40 hashtags (v1.12 — January 25, 2021)

Conservative/Trump campaign	Liberal/Biden campaign	Conspiracy	Other
trump2020	bidenharris2020	wwg1wga	coronavirus
trump	joebiden	stopthesteal	vote
kag	biden2020	qanon	election2020
americafirst	demconvention	dobbs	breaking
kag2020	demdebate		trumpvirus
maga2020	democrats		foxnews
trump2020landslide	yanggang		fakenews
			usa
			china
			georgia
			resist
			covid
			gapol
			fightback
			walkaway
			blacklivesmatter
			debates2020
			wethepeople

Table 4 Top 10 language breakdown for release v1.12. Languages were automatically tagged by Twitter and returned in a tweet's metadata

Language	ISO	# Tweets	Percentage
English	en	1,111,698,635	88.36%
Undefined	und	95,452,866	7.59%
Spanish	es	17,387,937	1.38%
French	fr	5,703,955	0.45%
Portuguese	pt	5,224,164	0.42%
Japanese	ja	3,368,223	0.27%
German	de	1,743,004	0.14%
Turkish	tr	1,700,836	0.14%
Indonesian	in	1,680,790	0.13%
Italian	it	1,585,394	0.13%

observed that there are highly connected conservative users that are more prone to spread public health and voting misinformation [2].

During the 2020 Presidential election, the incumbent former President Trump, faced little difficulty in securing the Republican nomination.[21] Although Trump did

[21] https://www.politico.com/story/2019/09/06/republicans-cancel-primaries-trump-challengers-1483126.

face three Republican challengers (Mark Sanford, Joe Walsh and Bill Weld), Trump earned 2395 delegate votes, an overwhelming majority.[22]

The Democratic primaries were more competitive, with a historic 28 candidates vying for the nomination.[23] However, as national poll results began to roll in and initial primary results were tallied, candidates began to drop out of the race (see Table 5 for dates candidates from both parties suspended their campaigns). The advent of COVID-19 in the United States in March 2020, and the ensuing regulations to encourage social distancing, forced the remaining campaigns to shift to a virtual models. The race narrowed down to two candidates: Vermont senator Bernie Sanders and former Vice President Joe Biden. As more primaries took place and results reported, it became clear that Biden would win the 1991 delegates needed to become the presumptive Democratic nominee[24]. Sanders conceded to Biden on April 8, 2020 and endorsed Biden.[25,26]

Overview of presidential candidate Twitter discourse

Our dataset specifically tracked 2020 US Presidential elections-related keywords and accounts. As a result, we expect to see that the captured discourse reflects major events that took place throughout our collection period. We limit our analysis to tweets from our dataset that were collected from January 2020 through December 2020.

The fight for the Democratic Presidential Nomination

We first investigate the chatter surrounding the Democratic primaries, as the race to win the nomination was competitive and multiple candidates emerged as favorites. While Biden may have held an early lead, Sanders, Elizabeth Warren and Pete Buttigieg were also serious contenders.[27] In Fig. 1, we tracked mentions of each of the Democratic presidential candidates' names and Twitter handles who were still campaigning in March 2020, and found the 7-day daily rolling average percentage of all collected tweets that mentioned each candidate. This particular time series ends on May 8, 2020, which is one month after Sanders conceded to Biden, and Biden became the presumptive Democratic presidential candidate.

Throughout the Democratic primary timeline in Fig. 1, we can see that the attention that specific candidates attract on Twitter fluctuates greatly. We can clearly see that Sanders and Warren initially led most of the discourse on Twitter in January 2020, but that Sanders would eventually dominate Twitter chatter throughout most of the primaries. This dominance continues until February 25, 2020, when James

[22] https://www.270towin.com/2020-republican-nomination/.
[23] https://www.politifact.com/article/2019/may/02/big-democratic-primary-field-what-need/.
[24] https://apnews.com/article/bb261be1a4ca285b9422b2f6b93d8d75.
[25] https://www.nytimes.com/interactive/2019/us/politics/2020-presidential-candidates.html.
[26] https://www.npr.org/2020/04/08/814291136/bernie-sanders-is-suspending-his-presidential-campaign.
[27] https://projects.fivethirtyeight.com/2020-primary-forecast/.

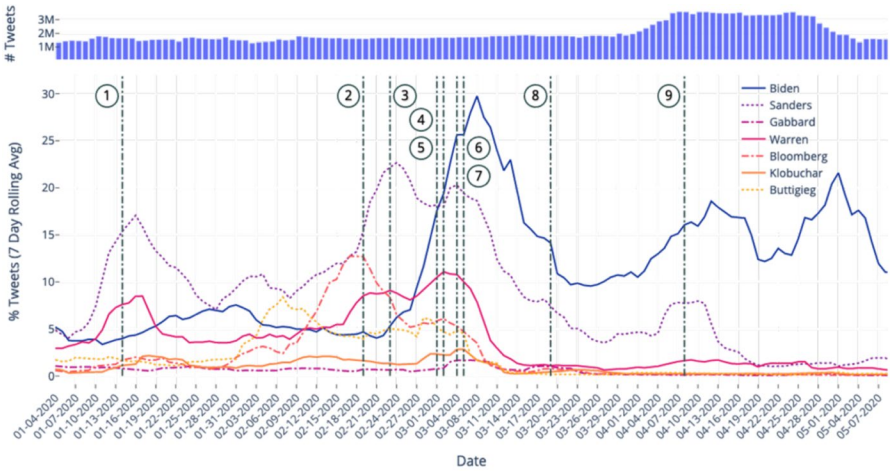
Table 5 This table lists each of the 2020 US Presidential candidates' names, party affiliation and campaign suspension date.

Candidate name	Party affiliation	Campaign suspended
Joseph R. Biden Jr.	Democrat	Democratic Nominee
Donald J. Trump	Republican	Republican Nominee
Bernie Sanders	Democrat	4/8/20
William F. Weld	Republican	3/18/20
Tulsi Gabbard	Democrat	3/19/20
Elizabeth Warren	Democrat	3/05/20
Michael R. Bloomberg	Democrat	3/04/20
Amy Klobuchar	Democrat	3/02/20
Pete Buttigieg	Democrat	3/01/20
Deval Patrick	Democrat	2/12/20
Andrew Yang	Democrat	2/11/20
Michael Bennet	Democrat	2/11/20
Joe Walsh	Republican	2/07/20
John Delaney	Democrat	1/31/20
Cory Booker	Democrat	1/13/20
Marianne Williamson	Democrat	1/10/20
Julin Castro	Democrat	1/02/20
Kamala Harris	Democrat	12/03/19
Steve Bullock	Democrat	12/02/19
Joe Sestak	Democrat	12/01/19
Wayne Messam	Democrat	11/20/19
Mark Sanford	Republican	11/12/19
Beto O'Rourke	Democrat	11/01/19
Tim Ryan	Democrat	10/24/19
Bill de Blasio	Democrat	9/20/19
Kirsten Gillibrand	Democrat	8/28/19
Seth Moulton	Democrat	8/23/19
Jay Inslee	Democrat	8/21/19
John Hickenlooper	Democrat	8/15/19
Eric Swalwell	Democrat	7/08/19
Richard Ojeda	Democrat	1/25/19

https://www.nytimes.com/interactive/2019/us/politics/2020-presidential-candidates.html

Clyburn, a prominent South Carolina African American Representative, endorsed Biden. From there, we see a sharp increase in Biden mentions, and Biden quickly overtook Sanders not only in polls, but also in Twitter discourse.[28] Biden continued to hold a majority in Twitter mentions throughout the rest of the primaries, through Sanders' concession on April 8, 2020. All other candidates saw a general decrease

[28] https://www.politico.com/news/2020/02/26/jim-clyburn-endorses-joe-biden-117667.



1	1/14/2020	7th Democratic Debate
2	2/19/2020	9th Democratic Debate; Bloomberg's debate debut received with much criticism.
3	2/23/2020	Sanders wins the Nevada Democratic caucuses.
4	3/01/2020	Buttigeg suspends his presidential campaign.
5	3/02/2020	Klobuchar suspends her presidential campaign.
6	3/04/2020	Bloomberg suspends his presidential campaign.
7	3/05/2020	Warren suspends her presidential campaign.
8	3/19/2020	Gabbard suspends her presidential campaign.
9	4/08/2020	Sanders suspends his presidential campaign.

Fig. 1 The above figure shows a time series analysis of tweets that mention keywords related to a Democratic nominee's campaign from January 2020 through May 8, 2020. Sanders announced the suspension of his presidential campaign on April 8, 2020, so we capture all discourse through a month after Biden was declared the presumptive Democratic Presidential nominee. We measure the percentage of total tweets collected on a particular day that mention the candidate on a rolling 7-day average. The keywords we use for each candidate can be found in Table 6 and descriptions of the noted dates in the table below the time series. We also include the raw volume of all tweets collected on a particular day on a rolling 7-day average above the time series

in tweet mention percentage after an initial increase in percentage after candidates announced that they had suspended their presidential campaigns.

While most of the mention percentages generally followed the popularity of certain candidates, in particular Biden, Sanders, Warren and Buttigieg, we find an increase in mentions surrounding Michael Bloomberg during the 9th Democratic debate.[29] The 9th Democratic debate was the first debate that Bloomberg was able to qualify for, but his performance was widely criticized.[30] He also attracted social

[29] https://www.pewresearch.org/fact-tank/2020/02/10/a-snapshot-of-the-top-2020-democratic-presidential-candidates-supporters/.

[30] https://www.npr.org/2020/02/20/807639778/6-takeaways-from-the-nevada-democratic-debate.

media attention after having heavily funded his campaign's ads with his personal money.[31]

Chatter during the Presidential elections: Biden versus Trump

We now turn to the final race in the 2020 U.S. Presidential election between Biden and Trump. As shown in Fig. 2 the percentage of all tweets that mention Trump is significantly greater than the percentage of tweets that mention Biden (see Table 6 for keywords associated with each candidate). This gap in mentions is not unexpected, as Trump was the incumbent President and thus already had a significant presence on Twitter. While our current analysis is based on percentage of mentions in the tweets collected, our prior work in clustering users by political affiliation based on shared media found that conservative users have a more vocal presence on the political Twitter scene [7]. Despite Trump's general dominance in the chatter, we see that as major events occur, such as when Democratic primaries began to be called for Biden and during the Presidential debates, Biden began to see an increase in mentions. While a tweet may be counted as mentioning both Trump and Biden, we still see a corresponding decrease in percentage of Trump's mentions when Biden's mentions increase. This suggests that the discourse shifted away from Trump and towards Biden, particularly as election day neared, culminating in a similar percentage of tweets mentioning either Biden and/or Trump.

It appears that the tweets we collected in our dataset track well the real world events. However, the sheer percentage of our collected tweets that mention a particular candidate does not necessarily represent the sentiment and popularity of those candidates at the time. As Twitter has evolved as a platform, likewise the user base has also changed [11]. This disparity between Twitter attention and real-world popularity was highlighted during the Democratic primaries. Sanders held the majority of percentage of tweet mentions from early January through the end of February. It was not until the initial primary results began to be tallied and reported that it became clear that Biden had actually won the Democrat's vote.[32] Sanders' dominance in Twitter discourse underscored how Biden's eventual momentum took much of the Democratic party by surprise.[33] This can give us insight into how news and public discourse on social media platforms can misrepresent or give a false impression of the nation's sentiment.

Twitter Location Engagement

Every tweet we collect is returned with metadata describing the tweet itself, including Twitter's automatic language tag and post date. Each tweet also includes

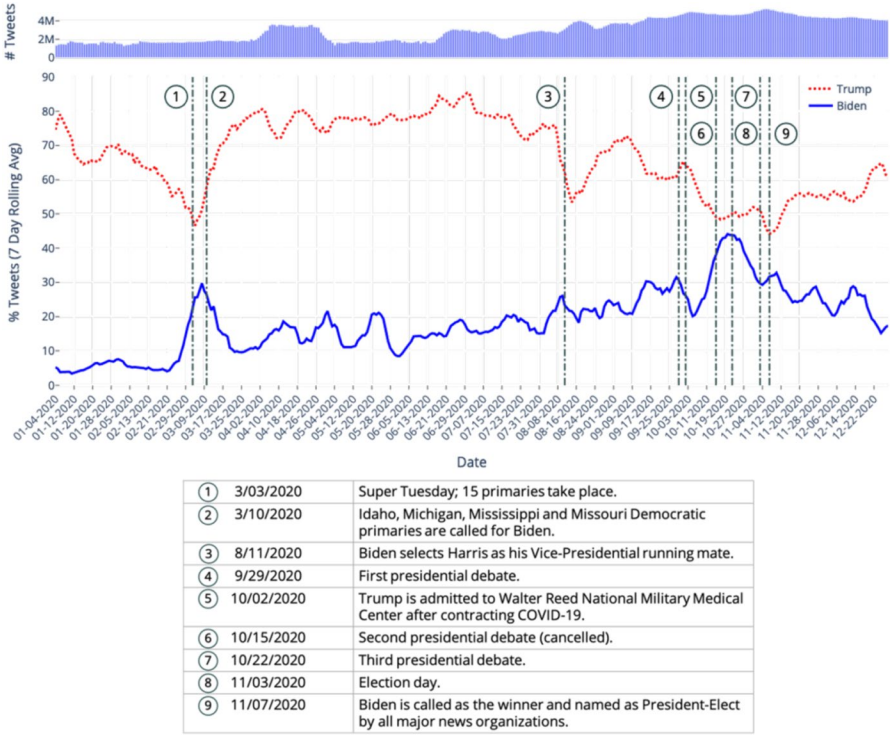
[31] https://www.npr.org/2020/02/21/808163144/bloomberg-has-already-spent-450-million-on-ads-since-launching-his-campaign.

[32] https://www.bbc.com/news/world-us-canada-52230979.

[33] https://www.npr.org/2020/03/04/811814716/bidens-surprise-win-in-texas-shows-momentum-may-matter-more-than-money.

Table 6 Keywords for each Democratic candidate that had not suspended their campaign by March 2020, and for Republican candidate Trump. We used these keywords to identify whether or not a candidate was mentioned in a tweet. We note that one tweet can be counted towards multiple candidates, if multiple candidates are mentioned in a tweet

Candidate name	Keywords
Donald J. Trump	@realDonaldTrump, realDonaldTrump, Donald Trump, DonaldTrump, Trump
Joseph R. Biden Jr.	@JoeBiden, JoeBiden, Joe Biden, Biden
Bernie Sanders	@BernieSanders, BernieSanders, Bernie Sanders, Sanders
Tulsi Gabbard	@TulsiGabbard, TulsiGabbard, Tulsi Gabbard, Gabbard
Elizabeth Warren	@ewarren, @senwarren, ewarren, senwarren, ElizabethWarren, Elizabeth Warren, Warren
Michael R. Bloomberg	@MikeBloomberg, MikeBloomberg, Mike Bloomberg, MichaelBloomberg, Michael Bloomberg, Bloomberg
Amy Klobuchar	@amyklobuchar, @senamyklobuchar, amyklobuchar, senamyklobuchar, amy klobuchar, klobuchar
Pete Buttigieg	@PeteButtigieg, PeteButtigieg, Pete Buttigieg, Buttigieg



1	3/03/2020	Super Tuesday; 15 primaries take place.
2	3/10/2020	Idaho, Michigan, Mississippi and Missouri Democratic primaries are called for Biden.
3	8/11/2020	Biden selects Harris as his Vice-Presidential running mate.
4	9/29/2020	First presidential debate.
5	10/02/2020	Trump is admitted to Walter Reed National Military Medical Center after contracting COVID-19.
6	10/15/2020	Second presidential debate (cancelled).
7	10/22/2020	Third presidential debate.
8	11/03/2020	Election day.
9	11/07/2020	Biden is called as the winner and named as President-Elect by all major news organizations.

Fig. 2 The above figure shows a time series analysis of tweets that mention keywords related to either Trump or Biden from December 2020 through January 2020. We measure the percentage of total tweets collected on a particular day that mention the candidate on a rolling 7-day average. The keywords we use for each candidate can be found in Table 6 and descriptions of the noted dates in the table below the time series. We also include the raw volume of all tweets collected on a particular day on a rolling 7-day average above the time series

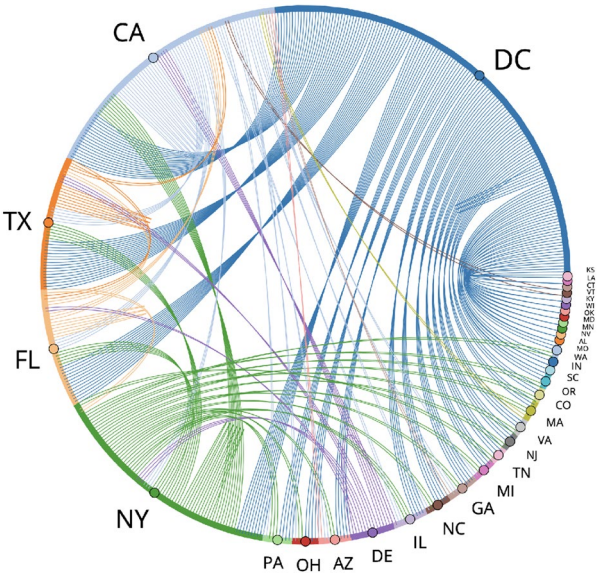


Fig. 3 We remove all tweets without an identifiable state, and visualize the intra-state tweet engagement activity within the United States in our dataset. For each retweet or quoted tweet, we visualize the geographic flow from the original poster's state to the retweeter's state. The line color corresponds to the location of the original poster

information about the author, and if the tweet was a response (reply, retweet or quote) to another tweet, the tweet's metadata also contains information on the original poster. This metadata can sometimes include a user's location data; however, we found that less than 1% of our tweets actually contained this information [9]. Because of this, we leverage the included "location" field that a user manually populates as a part of their profile. We tag each tweet with its country of origin and, if the tweet originates from the United States, the detected state [9]. While some users may list locations that are not accurate, do not exist or are unable to be identified through our algorithm, we leverage this as a proxy for tweet location.

We examine the domestic geographical flow of information within the United States. In isolating only retweets and quoted tweets (retweets with a comment), we find tweets that directly represent one user re-posting the tweet of another. Retweets and quoted tweets also return both the user specified location data for both the user who retweeted or quoted the tweet and the original poster. The user who retweeted or quoted the tweet will be referred to as the *retweeter* for clarity. Then, we retain all tweets within our dataset where we are able to identify a state for both the retweeter and the original poster, which directly implies that both the retweeter and original poster are also located in the United States. Figure 3 illustrates the flow of the top 200 most frequent state-to-state engagements, with the flow following retweets and quoted tweets from the original poster's state to the retweeter's state.

States in which the most tweets originate from generally coincide with the most populous states in the United States. The US Census Bureau lists California, Texas,

Florida and New York as the most populous states in their 2019 estimate.[34] However, most tweets actually originate from the District of Columbia area, which is both the political center and the capital of the United States. This is consistent with the nature of the political landscape, as many politicians are located in the D.C. area. In general, Fig. 3 suggests that while there exists a substantial amount of intra-state tweet engagement, states with larger populations account for larger proportions of the measured intra-state engagement activity.

Discussion

Limitations

While this dataset gives us a glimpse of the political chatter on Twitter, there are still limitations to this dataset that warrant discussion. Due to the nature of the keywords we were tracking, the tweets in our dataset are highly skewed towards English and tweets that originate from the United States. Another limitation of the dataset is that the users on Twitter do not necessarily represent the collective sentiment of the United States. The audience that uses Twitter, according to a 2019 study conducted by Pew Research Center, skews younger and more Democratic than the general population; the most vocal on Twitter also tend to engage in political discourse.[35]

Twitter also significantly rate limits the number of tweets that one can *rehydrate*, and tweets that have either been removed by the user or removed because a user was banned or suspended can no longer be retrieved through Twitter's API. Our collection was also highly contingent upon the stability of our network and hardware, which means that there may be gaps in our data collection, particularly prior to our migration to AWS. Twitter has recently released an Academic Research track that enables researchers and academics to access the full-archival search; however, this still imposes rate limits that unfortunately makes filling these gaps in time hard.[36]

Potential research avenues

There are many potential areas that can be explored using our dataset.

Recent work using our dataset has already begun to explore the prevalence of bots and misinformation within the 2020 political landscape [6, 7]. Luceri et al. also scrutinizes the bot engagement in political discourse in 2018 and found that many of these bots remained active during the 2020 election cycle [12]. Our previous work has found that out of all major conspiracy theories that had taken root during the election, QAnon supporters were the most vocal and active. We also found that, when grouping users by their political affiliation, tweets from accounts most likely to be bots outnumber tweets from accounts that are most likely human for both the Republican and Democratic parties. Conservative accounts that are the most likely

[34] https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html.
[35] https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.
[36] https://developer.twitter.com/en/solutions/academic-research.

to be bots also have higher bot scores, suggesting that these accounts are more likely to be automated compared to their left-leaning counterparts [7]. We used Indiana University's Botometer, a tool that assigns a bot-score to a Twitter account based on an account's activity [14, 15]. Others have also leveraged the polarized nature of the 2020 elections to model and estimate echo chambers based on a user's political stance [13].

While this is just a sampling of current literature, there are many areas that are also being explored, including the presence, effect and detection of trolls [8] and foreign influence during the elections [7]. Many new nascent and promising questions are also emerging in the wake of the elections, particularly as the COVID-19 pandemic has forced individuals to physically social distance and, consequently, seek community online.

After aggressive action to mitigate misinformation and the incitement of violence on major social network platforms, many flocked to alternative social network platforms that have espoused their support for freedom of speech, such as Parler and Gab.[37] While there has been much prior work in leveraging these alternative right-wing platforms to understand fringe views in conjunction with more main stream platforms [16–18] the recent high profile suspensions of major political figures' accounts led to an increased public awareness and exodus to these platforms. Before Parler went offline, researchers even scraped post data.[38] Data collected across multiple platform have the potential to give insight into how fringe communities not only survive these rebuffs by the community but also thrive in the controversy.

Another interesting question that arises is how the pandemic and the resulting shift to online platforms changed the nature and effectiveness of political campaigns. As some politicians quickly cancelled in-person events as the severity of COVID-19 rose, others chose to continue in-person rallies [1].[39,40] Social media became an integral part of the campaign process, more so than before, as events such as the Democratic National Convention were held virtually.[41] Cross-platform studies will be essential in beginning to understand the full scope of how and to what extent COVID-19 has fundamentally altered our elections system.

Conclusion

The 2020 US Presidential election cycle has been mired both by the COVID-19 pandemic and controversy. In this paper, we presented a Twitter dataset that we have collected from May 5, 2019 through the months after the transition to the Biden campaign. Twitter is by no means the only platform that campaigns leveraged to reach their base or where the public discussed their opinions. However, there has

[37] https://www.businessinsider.com/gab-reports-growth-in-the-midst-of-twitter-bans-2021-1.
[38] https://www.washingtonpost.com/technology/2021/01/12/parler-data-downloaded/.
[39] https://www.nytimes.com/2020/03/10/us/politics/sanders-biden-rally-coronavirus.html.
[40] https://www.cnn.com/2020/10/29/health/covid-trump-rallies-counties-cases/index.html.
[41] https://www.nytimes.com/2020/08/17/us/politics/democratic-national-convention-recap.html.

already been evidence that misinformation still persists on Twitter and other platforms, even as social media companies' are making efforts to address this problem [5–7]. Having access to this curated dataset will allow researchers to delve into how a contentious election unfolded and its surrounding chatter, as traditionally offline events transitioned online.

Inquiries

If you have technical questions about the data collection, please contact Emily Chen at https://www.echen920@usc.edu.

If you have any further questions about this dataset please contact Dr. Emilio Ferrara at https://www.emiliofe@usc.edu.

Acknowledgements The authors would like to thank Dr. Elizabeth Fife for her assistance in editing this manuscript.

Funding The authors gratefully acknowledge support from the Annenberg Foundation.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This data collection is based on public data and is registered as IRB exempt by the University of Southern California IRB (approved protocol UP-17-00610).

References

1. Bernheim, B. D., Buchmann, N., Freitas-Groff, Z., & Otero, S. (2020). *The effects of large group meetings on the spread of COVID-19: The case of Trump rallies*. USA: Stanford Institute for Economic Policy Research (SIEPR).
2. Chen, E., Chang, H., Rao, A., Lerman, K., Cowan, G., & Ferrara, E.(2021). Covid-19 misinformation and the 2020 u.s. presidential election. Special Issue on US Elections and Disinformation, Harvard Kennedy School Misinformation Review **1** . https://doi.org/10.37016/mr-2020-57
3. Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR, 6*(2), e19273.
4. Clayton, K., Blair, S., Busam, J., Forstner, S., Glance, J., Green, G., et al. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*. https://doi.org/10.1007/s11109-019-09533-0.
5. Ferrara, E. (2015). Manipulation and abuse on social media. *ACM SIGWEB Newsletter (Spring), 30,* 1–9.
6. Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*. https://doi.org/10.5210/fm.v25i6.10633.
7. Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020) . Characterizing social media manipulation in the, . (2020). US presidential election. First Monday. DOI 10.5210/fm.v25i11.11431. URL: https://journals.uic.edu/ojs/index.php/fm/article/view/11431
8. Jachim, P., Sharevski, F., & Pieroni, E. (2020). Trollhunter2020: Real-time detection of trolling narratives on twitter during the 2020 us elections
9. Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political polarization drives online conversations about COVID-19 in the United States. *HBET, 2,* 200–211.

10. Jungherr, A. (2016). Twitter use in election campaigns: a systematic literature review. *Journal of information technology and politics, 13*(1), 72–91. https://doi.org/10.1080/19331681.2015.1132401.
11. Liu, Y., Kliman-Silver, C., & Mislove, A.(2014). The tweets they are a-changin: Evolution of twitter users and behavior. Proceedings of the international AAAI conference on web and social media **8**(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14508
12. Luceri, L., Cardoso, F., & Giordano, S.(2020). Down the bot hole: actionable insights from a 1-year analysis of bots activity on twitter
13. Luo, R., Nettasinghe, B., & Krishnamurthy, V. (2020). Echo chambers and segregation in social networks: Markov bridge models and estimation
14. Sayyadiharikandeh, M., Varol, O., Yang, K.C., Flammini, A., & Menczer, F.(2020). Detection of novel social bots by ensembles of specialized classifiers. Proceedings of the 29th ACM international conference on information & knowledge management . Doi: 0.1145/3340531.3412698.
15. Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies, 1*(1), 48–61. https://doi.org/10.1002/hbe2.115.
16. Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., & Blackburn, J.(2018). What is gab: A bastion of free speech or an alt-right echo chamber. In: Companion proceedings of the the web conference 2018, WWW '18, p. 1007–1014. International world wide web conferences steering committee, republic and canton of Geneva, CHE . Doi: https://doi.org/10.1145/3184558.3191531.
17. Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J.(2019). Who let the trolls out? towards understanding state-sponsored trolls. In: Proceedings of the 10th ACM Conference on web science, WebSci '19, p. 353–362. Association for computing machinery, New York, NY, USA . Doi: https://doi.org/10.1145/3292522.3326016.
18. Zhou, Y., Dredze, M., Broniatowski, D. A., & Adler, W. D. (2019). Elites and foreign actors among the alt-right: The gab social media platform. *First Monday, 24*(9), 1. https://doi.org/10.5210/fm.v24i9.10062.