Methodology

This analysis examines a complete set of Facebook posts and tweets created on any account managed by any voting member of the U.S. Senate and House of Representatives during the months of February from 2015 through 2021. Researchers used the Facebook Graph API, CrowdTangle API and Twitter API to download the posts.[1] The resulting dataset contains 289,005 Facebook posts from 773 different members of Congress who used a total of 1,566 Facebook accounts, and 487,741 tweets from 775 different members of Congress who used a total of 1,595 Twitter accounts.

This analysis includes all text from these Facebook and Twitter posts, including image captions and emojis. Photo and video posts were not included in this analysis unless the post also contained meaningful text, such as a caption. Text that appeared only within images was not included in the analysis. Posts by nonvoting representatives were also excluded.

Here is more detail about the broader data collection process.

Identifying posts mentioning Black History Month

Researchers from Pew Research Center identified all posts over the entire time frame that mentioned Black History Month using a case-insensitive regular expression (a pattern of keywords and text formatting) matching the following key phrases and hashtags:

- "Black History"
- "Black History Month"
- "#BlackHistory"
- "#BlackHistoryMonth"

In total, 10,342 posts from the entire study period were identified as mentioning one or more of the keywords listed above and form the basis of this analysis.

Identifying the most common key words and phrases used in posts about Black History Month

Researchers also conducted a keyword analysis of the 10,342 posts identified as mentioning Black History Month. Text from each document (post) was converted into a set of features representing words and phrases by applying a series of pre-processing functions to the text of the posts. First, researchers removed 319 "stop words" that included common English words as well as the specific

[1] CrowdTangle is a public insights tool owned by Facebook.

phrase "Black History Month." The text of each post was then converted to lowercase, and URLs and links were removed using a regular expression. Punctuation was removed, and all terms were grouped into unigrams and bigrams (one- and two-word phrases). Researchers then ran a program to count the number of occurrences of each key word and phrase across all the posts mentioning Black History Month.

Identifying the most common historical figures mentioned in posts about Black History Month

In addition to identifying posts that mentioned Black History Month across the entirety of the study period, researchers from the Center conducted a detailed analysis of the posts made in February 2020 and February 2021 using a natural language processing technique called Named Entity Recognition (NER). NER is a form of information extraction where recognizable names (or *entities*) are identified in text and classified into predefined categories, such as geographic locations, the names of organizations, and the names of people. For example, a tweet from @NASA posted Feb. 1, 2022, reads:

> "Mae Jemison. George Carruthers. Katherine Johnson.
>
> This #BlackHistoryMonth, we're sharing stories of our many stars who light the way for future generations. Celebrate with us all month: https://go.nasa.gov/2GrOoU3"

Passing this text through a Named Entity Recognition model would identify "Mae Jemison," "George Carruthers" and "Katherine Johnson" as *named entities* belonging to the category *person*.

Center researchers used the NER pipeline in the Huggingface Transformers package for Python (Version 4.10.3) to process all 244,974 Facebook and Twitter posts created by members of Congress in February 2020 and February 2021, identifying all *person* entities referenced in these posts.[2] Then, coders manually reviewed the 500 most-referenced persons identified in the set of all posts, and the 500 most-referenced persons in the set of posts mentioning Black History Month, to identify cases where one individual is commonly referenced in more than one way. For instance, "Martin Luther King Jr." and "MLK" might be identified as separate entities, even though they refer to the same person. Where such cases were identified, coders unified these references into a

[2] By default, this function uses a version of the BERT-large Transformer language model that has been fine-tuned to perform Named Entity Recognition using the CoNLL-2003 dataset. More information about the NER pipeline can be found at https://huggingface.co/docs/transformers/task_summary#named-entity-recognition. The fine-tuned model we used can be found at https://huggingface.co/dbmdz/bert-large-cased-finetuned-conll03-english.

single, standardized representation. These unified references were then used to produce counts of how many members of Congress mentioned each individual figure.

The table below shows all references used to identify the 25 most-mentioned figures in posts about Black History Month from February 2020 and 2021.

Figure	Alternative references
Rosa Parks	'Rosa Parks', 'Rosa Park', 'RosaParks'
Martin Luther King Jr.	'King', 'Martin Luther King Jr', 'Martin Luther King', 'mlk', 'Martin Luther King Jr.', 'MLK Jr'
Kamala Harris	'Harris', 'Kamala Harris', 'Kamala', 'KamalaHarris'
Frederick Douglass	'Frederick Douglass', 'Douglass', 'Fredrick Douglass'
Katherine Johnson	'Katherine Johnson'
Shirley Chisholm	'Shirley Chisholm', 'Chisholm', 'ShirleyChisolm', 'ShirleyChisholm'
Hiram Revels	'Hiram Revels', 'Hiram Rhodes Revels', 'Hiram R. Revels', 'Revels'
Carter G. Woodson	'Carter G', 'Carter G. Woodson', 'Carter Woodson', 'Woodson'
John Lewis	'John Lewis', 'repjohnlewis'
Harriet Tubman	'Harriet Tubman', 'Tubman', 'HarrietTubman'
Barack Obama	'Obama', 'Barack Obama', 'BarackObama'
Maya Angelou	'Maya Angelou'
Claudette Colvin	'Claudette', 'Claudette Colvin', 'Colvin'
Clarence Thomas	'Clarence Thomas', 'Justice Clarence Thomas'
Joseph Rainey	'Joseph H. Rainey', 'Joseph Rainey'
Curt Flood	'Curt Flood'
Mary McLeod Bethune	'Mary McLeod', 'Mary McLeod Bethune'
James Baldwin	'James Baldwin'
Joyce Beatty	'Joyce Beatty', 'RepBeatty', 'JoyceBeatty'
Emmett Till	'Emmett Till', 'EmmettTill'
Tim Scott	'SenatorTimScott', 'Tim Scott'
Joe Biden	'Biden', 'Joe Biden', 'JoeBiden'
Madam C.J. Walker	'CJ Walker', 'J. Walker', 'Madame C. J. Walker'
Booker T. Washington	'Booker T Washington', 'Booker T. Washington'
Mae Jemison	'Mae Carol Jemison', 'Mae Jemison'