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 between Jan. 1, 2016, and April 30, 2021. Researchers used the Facebook Graph API, CrowdTangle API and Twitter API to download the posts. The resulting dataset contains more than 1.6 million Facebook posts from 773 different members of Congress who used a total of 1,520 Facebook accounts, and more than 3.5 million tweets from 772 different members of Congress who used a total of 1,466 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.

The broader data collection process is described in more detail here.

Identification of posts mentioning Asian countries and publics

Researchers from Pew Research Center identified all posts over the entire time frame that mentioned Asian countries and publics using a case-insensitive regular expression (a pattern of keywords and text formatting). The final set of keywords used to identify these posts included those listed below. Mentions of Asian individuals by name, outside the context of other relevant terms and keywords, were not coded for this analysis. Mentions of these terms immediately followed by "American" were also excluded, as were posts mentioning India or Indian that also contain terms referring to Native Americans (e.g., "Native American," "Indian Tribe," "Indian land," etc.).

Names of the most common Asian publics are drawn from the Asian American origin groups reported in the U.S. Census Bureau's American Community Survey:

- Bangladesh & Bangladeshi
- Bhutan & Bhutanese
- Burma (Myanmar) & Burmese
- Cambodia & Cambodian
- China & Chinese
- [The] Philippines & Filipino
- Hmong

- India & Indian
- Indonesia & Indonesian
- Japan & Japanese
- Korea & Korean (North Korea[n] & South Korea[n])
- Laos & Laotian
- Malaysia & Malaysian
- Mongolia & Mongolian
- Nepal & Nepalese
- Pakistan & Pakistani
- Sri Lanka & Sri Lankan
- Taiwan & Taiwanese
- Thailand & Thai
- Vietnam & Vietnamese

Posts mentioning of Korea or Koreans were further specified as mentioning South Korea, North Korea or both based on whether the posts mentioned terms that specifically refer to North Korea (e.g., "North Korea," "N. Korea," "Democratic People's Republic of Korea"), South Korea (e.g., "South Korea," "S. Korea," "Republic of Korea") or both Koreas (e.g., "Korean peninsula," "Korean War," "Korean (War) Veteran"). General mentions of "Korea" or "Korean" with no additional detail were classified as mentioning South Korea.

To evaluate the performance of the regular expression, researchers examined a random sample of 200 posts from Jan. 1, 2016, to April 30, 2021. Two researchers examined this set to determine whether they mentioned the relevant Asian country or public by name in order to compare human decisions with the decisions from the regular expression. Overall, the human decisions agreed with the keyword method 99% of the time. Cohen's Kappa was 0.98 for the same comparison. Another researcher also classified the set of 200 tweets independently to ensure their decisions were comparable. Cohen's Kappa for coder-to-code comparisons was 0.98.

In addition, researchers also evaluated the performance of regular expression in weeding out mentions of "India/Indian" that are not about the country and people of India by taking an additional sample of 150 posts that mentioned "India/Indian" before and after the regular expression pruning. In this case, human decisions agree with the keyword method 86% of the time, with Cohen Kappa of 0.8. Coder-to-coder Kappa is 1, or perfect agreement.

In total, 91,508 posts from the entire study period were flagged as mentioning one or more of the keywords listed above. They form the basis of this analysis.

Distinct keywords by party

Researchers also conducted a distinct keyword analysis using the complete set of 27,611 Facebook posts and tweets mentioning China created by members of Congress from Jan. 1, 2020, through April 30, 2021. The analysis was conducted on two separate sets of posts: those that mentioned the COVID-19 pandemic and those that did not.

Text from each document (post) was converted into a set of features representing words and phrases via a series of pre-processing functions to the text of the posts. First, researchers removed 3,132 "stop words" that included common English words, names and abbreviations for states and months, numerical terms like "first," and a handful of generic terms common on social media platforms like "Facebook" and "retweet." A set of stop words containing country names (e.g., "China," "Korea," etc.) and COVID-19 terms (e.g., "COVID," "pandemic," "coronavirus") were added to the general list of stop words for select relevant analysis. For distinct keyword analysis on China and the pandemic, researchers excluded posts that mentioned China only the context of a CDC link subtitle that describe COVID-19 as "a respiratory disease first detected in Wuhan, China." In total, 471 (6%) of these posts were identified and removed from the dataset of 8,044 posts that mentioned "China" or "Chinese" and the pandemic.

The text of each post was then converted to lowercase, and URLs and links were removed using a regular expression. Common contractions were expanded into their constituent words, punctuation was removed, and each sentence was tokenized using the resulting white space. Finally, words were lemmatized (reduced to their semantic root form) and filtered to those containing three or more characters. Terms were then grouped into two- and three-word phrases.

Distinctive keywords and phrases used by each party's members of Congress on each platform (Facebook and Twitter) were identified using pointwise mutual information. Researchers then calculated the proportion of party members who mentioned each distinct term (phrase). Terms mentioned by fewer than 10 members of either party are excluded. Researchers then used the proportions to calculate a ratio of differences in mentions between parties for each term. The most distinctive party keywords were defined as those terms with the largest ratio difference between the parties.

Finally, researchers consolidated phrases: removing those that had a word in common with any other phrase that was associated with a larger difference (e.g., "Chinese Communist Party" is not shown because "Chinese Communist" was associated with an even larger party difference).