Data Sets:

The s3mc project collected numerous data sets to support research related to the 2020 Election. The overall project goal is to understand the measurement properties of different types of organic data for understanding public opinion. We refer you to s3mc.org and smrconverge.org for more information about the project.

In this paper, we used 4 datasets:

Newspapers

We collected 186,551 articles from 308 newspapers around the country that contained the keywords "Biden" or "Trump" through EventRegistry. Using their API, we collected with any of our keywords ("Trump" or "Biden"). We cannot release these data for public use, but they can be used through the MDI Analytic Portal: portals.mdi.georgetown.edu.

Television

We obtained 24/7 closed caption transcripts from 1,246 television channels with news programming from TVEyes through their API. The data is broken down into 5 to 10 minute segments or chunks determined by advertising. We identified segments that referred to the US presidential election or their candidates using the words "election", "Trump", and "Biden". We cannot release these data for public use, but they can be used through the MDI Analytic Portal: portals.mdi.georgetown.edu.

Twitter

We used the Twitter API to collect tweets containing either the "Biden" or "Trump" keyword, yielding 62,343,263 and 107,410,289 tweets, respectively. We cannot release these data for public use, but they can be used through the MDI Analytic Portal: portals.mdi.georgetown.edu. We can also share the tweet id for researchers who want to use those to collect the associated tweets themselves.

Surveys

The CNN Breakthrough project collected 17,800 telephone surveys from July 1, 2020 to November 10 2020 among a nationwide, random sample of approximately 1,000 adults per week via the SSRS Omnibus survey. Surveys were conducted over a six-day period each week, typically between Tuesday and Sunday, in English and Spanish. Roughly 70% of surveys each week were completed with respondents reached via cellphone. Each respondent was asked, "What, if anything, have you heard, read or seen in the past few days about Donald Trump?" and "What, if anything, have you heard, read or seen in the past few days about Joe Biden?" The order in which the two questions were asked was randomized, so that some respondents were asked about Trump first and others were asked about Biden first. Exact responses to these questions were transcribed by interviewers. These data can also be accessed via the MDI Analytic Portal: portals.mdi.georgetown.edu.

Preprocessing:

When working with topics/myths across different data sets, preprocessing is needed to make the text more comparable and make the analysis more uniform. We used the following preprocessing steps:

Please feel free to use our preprocessing toolkit:
https://github.com/GU-DataLab/topic-modeling-textPrep

Preprocessing steps:

- Lowercased text
- Expanded contractions
- Removed punctuation

The analysis was run using an Apache Airflow Job on Google's Cloud Infrastructure. If you would like access to our code, please feel free to email any of the authors.

We would like to thank CNN and SSRS for sharing the open-ended responses through the CNN Breakthrough project. We would like to thank NSF for supporting this research.