# Data:

## Phase 1: Getting the Data from Reddit.

From Nov 22nd to Nov 24th, we collected 1000 post from each subreddit every day. (Totalling 6000 posts, 3000 from each subreddit). The default subreddit sort, hot, was used.

## Phase 2: Data Filtering (Automated).

The first filtering process was checking if Biden or Trump were mentioned in the post. we did that by checking for "Biden "or" Trump" (case insensitive) in the title and name of each post. Roughly 50% of all posts in r/politics (sample for liberals) and 20% of all posts in r/conservative (sample for conservative) mentioned either candidate. Overall, about 2100 posts remained after the first filter process.

Then, we removed all duplicates by checking the unique ID of every post. Overall, about 1200 posts remained after checking duplicates.

## Phase 2: Data Filtering (Manual).

We Removed the posts that mentioned "Trump" in reference to Donald Trump's family and not the candidate Donald Trump. Only a small number of posts were affected by the manual phase of the filtering process

## Phase 3: Building the dataset.

We reformatted the posts into a Tab Separated values text file with three columns: name, title and coding. We didn't include the "Self_ text" field because only a few posts in r/conservative included it and almost none of the r/politics posts did.

Overall, we had X Biden posts and Y Trump posts in r/politics. While r/conservative had X Biden posts and Y Trump posts.

These post-filtering posts include the following fields:

"name": unique ID for every post

"title": the title text.

"coding": for Manual coding.

An Example post from the dataset:

| name | title | coding |
|---|---|---|
| t3_jzpllg | The Already Useless Trump Kids Are About to Get Even More So | U |
| t3_jz1a2t | Trump Lawyer Sidney Powell Says Georgia Election Lawsuit 'Will Be Biblical,' Suggests GOP Governor Helped Biden | 1 |