Data Cleaning

Grace Abels

4/19/2022

Version 0 - All Unsorted Data

The dataset was pulled by Joel Luther. The dataset includes all entries from Claim Review that were published by sources based in the US. 14 unique fact-checking publishers were included. Included entries were those in which the reviewDate was between Jan 1st, 2016 and June 30th, 2021.

```
v0_all_us_unsorted <- read_csv("allUS_unsorted.csv")

v0_all_us_unsorted%>%
  count() %>%
  kable(caption = "All US Claims from ClaimReview 01/01/2016 - 06/30/2021")
```

Table 1: All US Claims from ClaimReview 01/01/2016 - 06/30/2021

n
18770

```
v0_all_us_unsorted %>%
  count(publisher.site)%>%
  kable(caption = "All US Claims from ClaimReview 01/01/2016 - 06/30/2021
        Sorted by Publisher")
```

Table 2: All US Claims from ClaimReview 01/01/2016 - 06/30/2021
Sorted by Publisher

publisher.site	n
cbsnews.com	240
checkyourfact.com	2108
factcheck.org	2310
factcheck.thedispatch.com	176
hoax-alert.leadstories.com	5
leadstories.com	3287
newsweek.com	189
nytimes.com	485
politifact.com	7770
polygraph.info	3
poynter.org	50
thegazette.com	9
usatoday.com	756
vox.com	2

publisher.site	n
washingtonpost.com	1380

NEED ASA TO EXPLAIN HOW TO GO FROM VERSION 0 to 1. Politician sort

Version 1 - Only Political Figures

"We built a small data pipeline that tried to identify, for each claim, whether its uterrer was a person and, if they were, whether they met our definition of a politician. A politician, we decided, was anyone who'd held or run for partisan office, or who'd been hired or appointed by a such a person to serve on a campaign or in a government agency.

We identified human claimants by feeding claimant names in our dataset through an entity recognition API. Given text like ""The Sierra Club"" or ""Bill Murray"", the API tried to detect what it referred to – e.g. a person or an organization – letting us label claimants as human.

Next, we tried to categorize human claimants as politicians or non-politicians. To do so, we fed claimants' names through a Wikipedia API. If a given claimant had a Wikipedia article about them, our code checked whether the article contained any ""politician indicators"". For example, if the infobox in an article about a claimant said they'd worked as a politician or political operative, we accepted that as fact and marked them as a politician. Likewise, if in the first few paragraphs of the Wikipedia article, text matched certain keywords (e.g. in a single sentence, the article mentioned its subject""ran"" in an ""election"" or ""work""ed in the""White House""), the code inferred that they, too, were a politician.

Additionally, once the code deduced that a claimant was a politician, it scanned their Wikipedia infobox for reference to their party affiliation, recording it in our dataset. If there was no Wikipedia article about a human claimant, as was occasionally the case for unsuccessful candidates for local office, the data pipeline tried to find a corresponding page for the claimant on Ballotpedia, an online encyclopedia of American politics. If it found one, it searched the page for the same markers looked for on Wikipedia pages, and similarly tried to identify the claimant's party affiliation. When the process above had been completed for each claimant, data collection was finished, and we filtered out all claims in the dataset whose claimants weren't labeled as human politicians. Following data collection, we undertook some light data cleaning. Namely, if we found that a claimant had been affiliated with several political parties, we gave them a party label corresponding to the party they'd identified with when they'd made the claim. In a small number of cases, our code categorized a claimant as a politician but failed to infer their party; those claimants' politician status and party ID were manually reviewed and edited at a later stage but were labelled as unknown_affiliation for the time being. The resulting dataset was of only claims made by politicians and political figures in the US with corresponding parties labeled.

Table 3: All Politician Claims (post-filter)

n
7635

Table 4: All Politician Claims (post-filter) Sorted by Publisher

publisher.site	n
cbsnews.com	167
checkyourfact.com	10
factcheck.org	1291
factcheck.thedispatch.com	59
leadstories.com	1

publisher.site	n
newsweek.com	66
nytimes.com	465
politifact.com	4282
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1253

Table 5: All Politician Claims (post-filter) Sorted by Party

claimant_party	n
[[Chinese Communist Party]]	1
[[Conservative Party (UK)|Conservative]]	1
[[Labour Party (UK)|Labour]]	2
[[Liberal Party of Canada|Liberal]]	1
[[Likud]]	2
[[Moderate Party|Moderate]]	1
[[UK Independence Party]] (2021–present)	1
[[Workers' Party of Korea]]	1
{{ubl|[[Peace and Freedom Party|Peace and Freedom]] (2012–2013)|[[Green Party of the United States|Green]] (2008–2012)}}	1
British Freedom Party (2018–present)	1
Democratic	2325
Independent	52
Libertarian	9
Republican	5119
unknown_affiliation	118

Version 2 - 5 Parties

At this stage we noticed that several of the claimant's were assigned non-US political parties suggesting they were not US political figures. At this point, all claims whose party affiliation in claimant_party was not Democratic, Republican, Libertarian, Independent, or unknown_affilition were removed. Version 2 only consists of claims made by political figures belonging to U.S. political parties or ones that we assigned an unknown affiliation. At this time, those with unknown_ affiliation were left unsorted. We addressed these once we had narrowed down to our smallest size datasets as to minimize hand sorting.

Table 6: All Claims in 5 Parties

n
7623

Table 7: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	167

publisher.site	n
checkyourfact.com	10
factcheck.org	1290
factcheck.thedispatch.com	58
newsweek.com	66
nytimes.com	464
politifact.com	4275
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1252

Table 8: All Claims in 5 Parties Sorted by Party

claimant_party	n
Democratic	2325
Independent	52
Libertarian	9
Republican	5119
unknown_affiliation	118

Version 3 - Deduped

At this step we filtered out duplicate claims. These were claims in which both the url and and text (previously thought title but it wasnt) of the claim were identical. The first appearance of each claim remained in the dataset, other were removed. This resulted in a database of 7,360 unique claims.

Table 9: All Claims in 5 Parties - Deduped

n
7360

Table 10: All Claims in 5 Parties - Deduped and Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	10
factcheck.org	1237
factcheck.thedispatch.com	58
newsweek.com	63
nytimes.com	444
politifact.com	4176
polygraph.info	2
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1

publisher.site	n
washingtonpost.com	1179

Table 11: All Claims in 5 Parties - Deduped and Sorted by Party

claimant_party	n
Democratic	2265
Independent	50
Libertarian	9
Republican	4921
unknown_affiliation	115

Version 4 - Dejunked

For a final clean up before sorting by publisher and removing Trump, a member of our research team **manually identified 23 false positive rows** that either listed multiple claimants or had a false positive non-human claimant (e.g. "Donald Trump's campaign"). Those 23 claims were manually removed.

Table 12: All Claims in 5 Parties - Deduped and Dejunked

n
7337

Table 13: All Claims in 5 Parties - Deduped and Dejunked Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	9
factcheck.org	1237
factcheck.thedispatch.com	53
newsweek.com	62
nytimes.com	444
politifact.com	4167
polygraph.info	2
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1172

Table 14: All Claims in 5 Parties - Deduped and Dejunked Sorted by Party

claimant_party	n
Democratic	2259
Independent	50

claimant_party	n
Libertarian	9
Republican	4907
unknown_affiliation	112

At this stage, the dataset of 7,337 claims had been cleaned of irrelevant and redundant data removed. The dataset that remains consists of claims that were made by U.S. political figures, and fact-checked by one of 13 U.S based publishers. At this stage, we began working with each publisher dataset separately. Only four publishers at this stage had enough claims to move forward: FactCheck.Org, The New Your Times, PolitiFact, and The Washington Post. Go to the next sheet for our next step.

Version 5 - Split into Publisher data

At this point we began looking at the makeup of each publisher individually. Based on the number of total fact checks, we selected four publishers to analyze in the next step: PolitiFact, The Washington Post, FactCheck.Org, and The New York Times. This resulted in four individual datasets, one per publisher, each denoted as v5.

Table 15: Politifact - All Claims in 5 Parties

n
4167

Table 16: Politifact - All Claims in 5 Parties - Sorted by Party

claimant_party	n
Democratic	1529
Independent	23
Libertarian	8
Republican	2535
unknown_affiliation	72

Table 17: Washington Post - All Claims in 5 Parties

n
1172

Table 18: Washington Post - All Claims in 5 Parties - Sorted by Party

claimant_party	n
Democratic	293
Independent	12
Republican	855
unknown_affiliation	12

Table 19: FactCheck.Org - All Claims in 5 Parties

n
1237

Table 20: FactCheck.Org - All Claims in 5 Parties - Sorted by Party

claimant_party	n
Democratic	288
Independent	12
Libertarian	1
Republican	922
unknown_affiliation	14

Table 21: New York Times - All Claims in 5 Parties

n
444

Table 22: New York Times - All Claims in 5 Parties - Sorted by Party

claimant_party	n
Democratic	101
Republican	341
unknown_affiliation	2

Version 6 - Removing Trump

We made the decision to remove Donald Trump entirely from this dataset. To do so, we first identified all the different iterations of his name within the `claimant` column. We could not simply remove all claimants whose names included Trump, because under our definitions, his daughter Ivanka Trump was a politician. The code determining these iterations is shown below.

Table 23: Names including 'Trump'

claimant	n
Donald J. Trump	205
Donald trump	1
Donald Trump	2587
Ivanka Trump	10
President Donald J. Trump	2

With these four versions of Trump's name, we manually removed them from each V5 dataset. This created our V6 dataset without Trump for each publisher.

Table 24: Politifact - Without Trump

n
2951

Table 25: Politifact - Without Trump - Sorted by Party

claimant_party	n
Democratic	1529
Independent	23
Libertarian	8
Republican	1319
unknown_affiliation	72

Table 26: Washington Post - Without Trump

n
568

Table 27: Washington Post - Without Trump - Sorted by Party

claimant_party	n
Democratic	293
Independent	12
Republican	251
unknown_affiliation	12

Table 28: FactCheck.Org - Without Trump

n
526

Table 29: FactCheck.Org - Without Trump - Sorted by Party

claimant_party	n
Democratic	288
Independent	12
Libertarian	1
Republican	211
unknown_affiliation	14

Table 30: New York Times - Without Trump

n
180

Table 31: New York Times - Without Trump -Sorted by Party

claimant_party	n
Democratic	101
Republican	77
unknown_affiliation	2

Version 7 - Sorting Unknown Affiliation

After we had the data divided into its smaller publisher sub-sets, and removed Trump from the data, our next step was to address anomalies in the claimant_party section. The code that was used to assign party affiliations scraped Wikipedia data and Ballotpedia for politicians and party affiliations. For about 100 different claims, the claimant was identified as a political figure but a party could not be found. They were labeled unknown_affiliation. We did not make an attempt to sort these previously, because there were simply too many. However, at this stage we have narrowed down to the data we plan to use. Therefore our next step was to manually assign party affiliations to the unknown_affiliation claims.

This was done by pulling the claims labeled unknown_affiliation for each publisher. Only the fields using the name of the claimant, and occasionally the url for the fact-check itself if ambiguous, claimants were either

1. Assigned a party (Democrat, Republican, Independent, Libertarian)
2. Marked to be removed from the dataset because they were not a political figure.

For all those assigned a party, the cell to the left denotes why and provides a link to the source used to make that determination. If someone changed parties, the claim date was noted.

NOTE: If a claimant served in an administration and spoke on behalf of that administration or politician, they were considered a political figure and assigned the affiliation of the administration or politician they served/spoke on behalf of. This applied to appointed cabinet members, political officials, and lawyers representing politicians. Political commentators, activists, and celebrities were removed and not considered political figures