Data Cleaning

Grace Abels

4/19/2022

```
library(tidyverse)
library(knitr)
```

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

publisher.site	n
thegazette.com	9
usatoday.com	756
vox.com	2
washingtonpost.com	1380

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.

```
v1_all_us_polclaims <- read_csv("all_us_pol_claims_updated_sep15.csv")

v1_all_us_polclaims %>%
  count() %>%
  kable(caption = "All Politician Claims (post-filter)")
```

Table 3: All Politician Claims (post-filter)

n
7635

```
v1_all_us_polclaims %>%
  count(publisher.site) %>%
```

```
  kable(caption = "All Politician Claims (post-filter)
        Sorted by Publisher")
```

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
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

```
v1_all_us_polclaims %>%
 count(claimant_party) %>%
 kable(caption = "All Politician Claims (post-filter)
        Sorted by Party")
```

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.

```
v2_5parties <- v1_all_us_polclaims %>%
  filter(claimant_party == "Republican"|
           claimant_party == "Democratic"|
         claimant_party == "Independent"|
         claimant_party == "Libertarian"|
         claimant_party == "unknown_affiliation")

v2_5parties %>%
  count() %>%
  kable(caption = "All Claims in 5 Parties")
```

Table 6: All Claims in 5 Parties

n
7623

```
v2_5parties %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 7: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	167
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

```
v2_5parties %>%
  count(claimant_party) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Party")
```

Table 8: All Claims in 5 Parties Sorted by Party

claimant_party	n
Democratic	2325

claimant_party	n
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.

```
v3_deduped <- v2_5parties %>%
  distinct(url, text, .keep_all = TRUE)
```

```
v3_deduped %>%
  count() %>%
  kable(caption = "All Claims in 5 Parties - Deduped")
```

Table 9: All Claims in 5 Parties - Deduped

n
7360

```
v3_deduped %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties - Deduped and
        Sorted by Publisher")
```

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
washingtonpost.com	1179

```
v3_deduped %>%
  count(claimant_party) %>%
  kable(caption = "All Claims in 5 Parties - Deduped and
        Sorted by Party")
```

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.

```
v4_dejunked <- v3_deduped %>%
  filter(claimant != "Donald Trump For Prison") %>%
  filter(claimant !=  "Lauren Boebert; Rudy Giuliani") %>%
  filter(claimant != "Rick Scott's Starbuck's heckler") %>%
  filter(claimant != "Americans United for Change") %>%
  filter(claimant !=  "Consumers for Smart Solar") %>%
  filter(claimant !=  "Greg Gianforte's campaign") %>%
  filter(claimant !=  "Vietnam Veterans Against John McCain") %>%
  filter(claimant != "President Trump's lawyers") %>%
  filter(claimant !=    "Robin Vos and Scott Fitzgerald") %>%
  filter(claimant !=  "Robin Vos; Scott Fitzgerald") %>%
  filter(claimant !=  "Donald Trump 2020 Voters") %>%
  filter(claimant !=  "Michael Bloomberg; Joe Biden; Hillary Clinton; Adam Schiff") %>%
  filter(claimant !=  "Donald Trump and Mike Pence") %>%
  filter(claimant !=  "Bill DeBlasio and Brian Kemp") %>%
  filter(claimant !=  "John Roberts; Fox News correspondent") %>%
  filter(claimant !=  "The Trump campaign") %>%
  filter(claimant !=  "Sen. Ted Cruz (R-Texas) and Rep. Mark Meadows (R-N.C.)") %>%
  filter(claimant !=  "Keith Ellison spokesman") %>%
  filter(claimant !=  "Donald Trump ad") %>%
  filter(claimant !=  "Kamala Harris for the People") %>%
  filter(claimant !=  "Don Bolduc campaign") %>%
  filter(claimant !=  "Donald Trump campaign") %>%
  filter(claimant !=  "Joe Biden campaign")
```

```
v4_dejunked %>%
  count() %>%
  kable(caption = "All Claims in 5 Parties - Deduped and Dejunked")
```

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

n
7337

```
v4_dejunked%>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties - Deduped and Dejunked
        Sorted by Publisher")
```

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

```
v4_dejunked %>%
  count(claimant_party) %>%
  kable(caption = "All Claims in 5 Parties - Deduped and Dejunked
        Sorted by Party")
```

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

claimant_party	n
Democratic	2259
Independent	50
Libertarian	9
Republican	4907
unknown_affiliation	112

Version 5 - Split into Publisher data

```
pf_v5 <- v4_dejunked %>%
  filter(publisher.site == "politifact.com")

wapo_v5 <- v4_dejunked %>%
  filter(publisher.site == "washingtonpost.com")

fc_v5 <- v4_dejunked %>%
  filter(publisher.site == "factcheck.org")

nyt_v5 <- v4_dejunked %>%
  filter(publisher.site == "nytimes.com")
```