PolitiFact Clean Up

Grace Abels

4/19/2022

```
v8_pf <- read_csv("v8pf.csv")
```

```
## Rows: 2881 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): url, title, textualRating, languageCode, publisher.name, publishe...
## dbl   (1): ...1
## dttm  (2): reviewDate, claimDate
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Version 8 - Removal of PolitiFact anomalous claims

"At this part of the process, we have sorted out all claims that do not fit our criteria of claims by U.S politician from out 4 publisher datasets. We have also eliminated Independent and Libertarian claims. Now within each publisher we zoom into the field textualRating. This denotes the rating that the claim was given by the fact-checker. Some have standardized scales with ranking others do not. For this step we have defined the existing rating scale, seen how many fit into that scale, and how many anomaly claims remain.

For PolitiFact, they have an established 6 tier rating scale called the Truth-O-Meter in which claims are assigned one of the following ratings. The scale goes from more true to less true. This information was pulled from PolitiFact's website

"The goal of the Truth-O-Meter is to reflect the relative accuracy of a statement. The meter has six ratings, in decreasing level of truthfulness:"

TRUE – The statement is accurate and there's nothing significant missing.

MOSTLY TRUE – The statement is accurate but needs clarification or additional information.

HALF TRUE – The statement is partially accurate but leaves out important details or takes things out of context.

MOSTLY FALSE – The statement contains an element of truth but ignores critical facts that would give a different impression.

FALSE – The statement is not accurate.

PANTS ON FIRE – The statement is not accurate and makes a ridiculous claim.

The burden of proof is on the speaker, and we rate statements based on the information known at the time the statement is made.

Below is a breakdown of all the ratings in PF textualRating. There are

2,527 Normal claims (those rated with on of the 6 standardized ratings)

354 Anomalous claims (those with a non-standardized text based rating)

Table 1: Textual Ratings

textualRating	n
Half True	555
Mostly True	534
Mostly False	499
FALSE	477
TRUE	329
Pants on Fire	133
Needs context	19
Full Flop	18
Accurate	14
Misleading	9
Half Flip	8
Largely correct	5
This is accurate	4
Correct	3
Lacks context	3
Largely accurate	3
Needs more context	3
Not the full story	3
Close to accurate	2
Exaggerated	2
Exaggerates	2
No evidence	2
No Flip	2
Not full story	2
This checks out	2
This is correct.	2
Wrong	2
‘Slip of the tongue’	1
A fair summary	1
Academic study backs it up	1
Accurate as to days	1
Ad watch	1
Admits it’s not true	1
Affects a small group	1
Aimed at Bernie Sanders; and it’s True.	1
Also didn’t say no	1
An element of truth; but exaggerated and misleading	1
Analysis (no rating given)	1
Analyzing president plan Bs	1
Bashed Obama’s foreign policy	1
Basically correct	1
Below other figures	1
Best we can tell	1
Biden opposes defunding police	1
Biden; others are named	1

textualRating	n
Bill was not a ‘game-changer’	1
Blends three heroic events	1
But his staff does	1
Capital flub	1
Cherry-picked time frame	1
Children may be last	1
China was huge factor	1
Cites early Trump remarks	1
Claim has notable support	1
Close; unclear meaning	1
Congress can set rules	1
Correct about grand juries	1
Correct on 2016; Trump	1
Correct; for Canada	1
Correct.	1
Courts may allow it	1
D-minus in 2012 from NRA’s political action committee	1
Data backs it up	1
Data checks out	1
Data from one survey	1
Deadline lacks nuance	1
Depends how you count	1
Depends on intent	1
Depends on source	1
Depends on the year	1
DNC didn’t direct	1
Don’t count your chickens	1
Downplays the figure	1
Earlier scores were weak	1
Economy was already growing	1
Editing amps up effect	1
Effect not clear	1
Employment and business losses have been large	1
Endorsed Obama in 2012; but not in 2008	1
Estimated to pay for 2019; not for 2017 and 2018	1
EU; not Ukraine	1
Event with Andrew Cuomo	1
Evidence is growing	1
Exaggeration	1
Experts say concept is sound	1
Experts say it’s credible	1
Experts: Study is narrow	1
Fair summary	1
Filibuster; racism intertwined	1
First wave not over for many	1
Floated having someone run	1
Follows 1946 law	1
Foreign ties are unusual	1
Found guilty by jury	1
Frequent attack needs context	1
Future uncertain	1
Generally accurate	1

textualRating	n
GOP never kicked out	1
Had seat at table	1
Handled by lower-level attorneys	1
Hard to confirm	1
Hasn't said why	1
He advised Trump campaign	1
He can't; experts say	1
He gave to GOPers	1
He heard from Mueller	1
Heavily disputed	1
High estimate	1
Hunter lacked expertise	1
Ignores Biden's current stance	1
Ignores big spending bills	1
Ignores military aid given	1
Ignores new distribution plan	1
In the ballpark	1
Inaccurate	1
It depends	1
It helps and hurts	1
Judge rejected bias claims	1
Key checks continue	1
Key metrics back it up	1
Lacks solid numbers	1
Largely right; needs details	1
Law doesn't say that	1
Leaves out key details	1
Legal experts agreed	1
Legal experts divided	1
Legal for Riot Control	1
Legal rules unclear	1
Less than totally candid	1
Lincoln's motivations unclear	1
Link goes back before Civil War	1
Linkage isn't perfect	1
Loaded language confuses issue	1
Lots more than that	1
Lots of context needed	1
Majority are adults	1
Many experts would disagree	1
Many guns still allowed	1
Many wanted Shokin fired	1
Matches CDC statements	1
Metering affects border crossers	1
Minor issues so far	1
Misleading on two counts	1
Misleads in several ways	1
Misrepresents his remarks	1
Misses a key point	1
Misses the full story	1
Missing context	1
More adding paper based	1

textualRating	n
More like three	1
More like true estimates	1
More words than action	1
Most not considered permanent	1
Mostly right	1
Mueller probe not baseless	1
Multiple figures confirm it	1
Narrowly accurate	1
Nearly 400;000 in FY2018	1
Needs full context	1
New data says so	1
News accounts confirm it	1
No contact	1
No direct link	1
No he didn't	1
No legal status granted	1
No proof of fraud	1
Not conclusive	1
Not everyone is gaining	1
Not fact witness' job	1
Not fully accurate	1
Not locked out	1
Not many; but some	1
Not provable	1
Not proven	1
Not settled among experts	1
Not settled science	1
Not the best example	1
Not the only reading	1
not the only reason	1
Not the whole story	1
Not What Volker Said	1
Numbers miss key details	1
Obama Called in 2012	1
Offered meeting; military aid	1
Omits key caveats	1
Only for reported crimes	1
Only true for men	1
Open legal question	1
Opposed war after start	1
Others also got cuts	1
Outdated number	1
Pandemic not in rear-view mirror	1
Partially accurate	1
Plausible pledge; experts say	1
Politico offers no support	1
Possible; but no legal slam dunk	1
Probe had wider scope	1
Proposed cuts were reversed	1
Puerto Rico loses most	1
Quickly corrected slip-up	1
Reality is complex	1

textualRating	n
Record is mixed	1
Reflects what reports say	1
Reports heavily disputed	1
Revenues might fall short	1
Right on jobs	1
Roughly right	1
Said someone should try	1
Schiff's evidence is weak	1
Selective examples	1
Shaky details	1
She has	1
She told law schools	1
Slow gains; but gains	1
Some evidence to support	1
Some reduced staff	1
Sondland cites names	1
Spinning who gets credit	1
State Police were prepared	1
Sterling presented multiple pieces of evidence to refute President Donald Trump's claims	1
Strain; but checks continue	1
Study backs it up	1
Stunt ad makes false claim	1
Suspicions; no final proof	1
Tests sent; not given	1
Thanks to 'extra' doses in vial	1
That's part of the story	1
That's wrong	1
The accuracy is mixed	1
The law is broad	1
They didn't decertify.	1
This is accurate.	1
This is exaggerated.	1
This is true	1
This is True	1
This is unproven	1
This is wrong	1
This needs context	1
This needs context.	1
Timeline doesn't match	1
Tough promise to keep	1
True based on available data	1
True but didn't coin	1
True for total vote counts	1
True; but needs context	1
Trump has raised idea	1
Trump role emerged later	1
Trump shifts stances	1
Trump tweeted about Trudeau	1
Trump was wrong	1
Twisted history	1
U.S. depends on Russia	1
U.S. does fare worse	1

textualRating	n
U.S. ranks near bottom	1
Unable to verify	1
Unclear	1
Unproven speculation; no evidence	1
Unproven with state data	1
Up; not that much	1
Virus now only trails Civil War in deaths	1
Wages have risen modestly	1
We track ethics reversal	1
Wine cave event occurred	1
Words are in booklet	1
Would face legal challenges	1
Wrong meeting	1
Wrong wording	1
Wrong; he mispoke	1
Yes; but general comments	1
Yes; nine schools cut	1
Yes; they flip- flopped	1

When we reviewed the list of anomaly claims, we noticed that all came from PolitiFact articles. PolitiFact values their Truth-O-Meter ratings and use them only when they feel they can convey some level of certainty in the rating given. Truth-O-Meter ratings require a high threshold of proof. When that is lacking or there is not evidence to do a full scale fact-check the existing facts are published in an article. Fact-checks of debates and speeches are frequently written up in articles. Since PF did not feel comfortable enough to deliver a full Truth-O-Meter rating neither did we and all claims that were not given on the 6 standard ratings in textualRating were removed from the dataset.

Version 9 - Standardized Ratings Only

Table 2: PF Claims with Truth-O-Meter Ratings

textualRating	n
Pants on Fire	133
False	477
Mostly False	499
Half True	555
Mostly True	534
True	329

After this removal we began looking closely at the remaining claimant names. Version 8 had 723 unique claimant names.

```
## # A tibble: 724 x 2
## # Groups:   claimant [724]
##    claimant            n
##    <chr>           <int>
##  1 Joe Biden         113
##  2 Hillary Clinton   104
##  3 Bernie Sanders     58
```

```
##  4 Mike Pence         56
##  5 Newt Gingrich      45
##  6 Ted Cruz           42
##  7 Marco Rubio        39
##  8 Ron Johnson        32
##  9 Barack Obama       31
## 10 Andrew Cuomo       30
## # ... with 714 more rows
```

Version 10 - Condensing claimant names

At this stage we noticed that some claimants were listed under several separate names referring to the same person. Like Speaker Nancy Pelosi, Nancy Pelosi, Speaker Pelosi. We wanted to eliminate this repetition so that we could see the true # of claims made by each claimant. To do so, we moved our data into a program called OpenRefine. Here, we clustered claimant names by similar terms to identify where repetition/multiple names for the same person occured. We used this to identify all duplicate forms of name and then recoded the data accordingly.

This processed combined several names making the list 25 names shorter. 698 unique claimants remain.

```
## # A tibble: 697 x 2
## # Groups:   claimant [697]
##    claimant            n
##    <chr>           <int>
##  1 Joe Biden         113
##  2 Hillary Clinton   104
##  3 Bernie Sanders     58
##  4 Mike Pence         56
##  5 Newt Gingrich      45
##  6 Ted Cruz           42
##  7 Marco Rubio        39
##  8 Ron Johnson        32
##  9 Andrew Cuomo       31
## 10 Barack Obama       31
## # ... with 687 more rows
```

Version 11 – Final Claimant Cleaning

During this process we also noticed that some claimants, who did not fit our defintion of politician, had slipped through the cracks in our code. To try and ensure that we had only the data we desired in our dataset, we ran the list of claimant names through a stricter version of the politician filter. 58 names were marked as potentially non-politicial figures. Each name was manually reviewed and we identified 6 names that did not belong in the dataset.

Tucker Carlson, Laura Ingraham, Jacob Wohl, State representatives, Reagan was Right, Marco Rubio's heckler. 21 claims were removed as a result.

Later on during tagging we identified three more claimants (Pat Robertson, Juan Williams, and Evan Smith) and 1 claim that was mislabeled (The claim said it was Maxine Waters but the link said it was bloggers) that were not political figures. For ease we have removed them here. 6 more claims were removed.

Below are the counts for the final dataset used for tagging.

Table 3: Claim Rating by Party for Final Dataset

textualRating	Democratic	Republican
Pants on Fire	31	93
False	160	310
Mostly False	218	273
Half True	316	237
Mostly True	362	171
True	218	111

In this process an erroneous claim came to our attention, dated 2106 instead of 2016. We manually recoded this.

Tagged Claims

HOW WE TAGGED AND WHY

WHY WE JOINED BY JUST THESE VARIABLES

Manual Removal of Duplicate Claims

EXPLAIN DUPLICATE REMOVAL

Table 4: Remaining Duplicates Identified during Tagging

text	n
We put a lid on Iran's nuclear program without firing a single shot.'	4
Illegal immigration and the crime rate are as low as they've been in decades.'	3
It was Hillary Clinton who left Americans in harm's way in Benghazi and after four Americans fell said; 'What difference at this point does it make?' '	3
Neighborhoods have become more violent' under President Barack Obama's 'watch.'	3
Says 'Donald Trump has defended' World War II internment camps.	3
Says Donald Trump 'claimed our armed forces are 'a disaster."	3
Says Hillary Clinton 'said all work-related emails were sent back to the State Department. The FBI director said; that's not true.'	3
Says; regarding the presence of classified information in her email; FBI Director James 'Comey said my answers were truthful; and what I've said is consistent with what I have told the American people.'	3
Trump ties are made 'in China; not Colorado. Trump suits in Mexico; not Michigan. Trump furniture in Turkey; not Ohio. Trump picture frames in India; not Wisconsin.'	3
'What difference; at this point; does it make?' I am the guy that got under her skin and provoked that infamous response from Hillary Clinton by asking a pretty simple question; 'Why didn't you just pick up the phone and call the survivors' (of the Benghazi attack)?'	2
John McCain's chief economic adviser during the '08 race . . . estimated that Trump's promises would cause America to lose 3.5 million jobs.'	2
Says '(Clinton) called President Assad a 'reformer.' She called Assad a 'different kind of leader.' '	2
Says 'Hillary (Clinton) wants to increase the number (of Syrian refugees) by 500 percent.'	2
Says as Indiana governor; he has made 'record investments in education.'	2
Says Donald Trump 'cashed in' on Sept. 11; 'collecting $150;000 in federal funds intended to help small businesses recover — even though days after the attack Trump said his properties were not affected.'	2
Says Hillary Clinton 'abided by the ethics agreement' between the Clinton Foundation and the Obama administration.	2
Says Hillary Clinton 'has been a champion of globalist trade agreements. . . . Worst of all; they are now pushing the disastrous 5;000-page Obamatrade — the Trans-Pacific Partnership agreement.'	2

text	n
Says Jim Sensenbrenner 'has been in office for 40 years' and 'he's led on exactly one bill;' the Patriot Act.	2
Says unlike Tim Kaine; who 'invested' in education; Indiana Gov. Mike Pence 'slashed education funding.'	2
The top one-tenth of 1 percent now owns almost as much wealth as the bottom 90 percent.'	2
We moved 100 times as many people out of poverty as moved out when President (Ronald) Reagan was in office; with 40 percent more jobs.'	2

One of the things that we tagged by was location of the lie, meaning the medium and format of the statement containing the lie. This tracked things like whether it was said on TV, in an interview, on social media, etc. We had several overarching categories called `location` tags, and then more specific subcategories tagged `location.extra`. When loading this data into R, we made a select number of changes to make it compatible with the existing data. We matched the name of the claim identifier column in the location of lie data to be identical to the mega data and selected only half the variables leaving only new information and the variables needed to join the two datasets correctly. We also found that one of the variables had not loaded in correctly, so we informed R that the `claimDate` column was, in fact, containing time/date data and fixed one erroneous claim where the year was misstyped 2106 instead of 2016.

We selected to join the location of lie data to the larger dataset by the claim identifier column, `...1`, and by `url`, `languageCode`, `publisher.name`, `publisher.site`, `text`, `claimDate`, and `claimant_party`. We chose not to merge by title or text of claim due to a small number of claims that were encoded oddly with incorrect symbols. This error likely occured during exportation and importation of the dataset. The correct title and text were pulled from the megadata, and we are certain these are still joined correctly thanks to the claim identifier column. We found one more incorrect claim that was missing the correct text, we sourced the accurate claim from the URL and overwrote it in our data. Then, we exported the final dataset into a viewable and downloadable csv, called `pf_mega_location.csv`.

```
pf_mega_location <- left_join(pf_mega_nodupes, loc_of_lie, by = c("...1", "url", "languageCode", "publis
pf_mega_location$title <- pf_mega_location$title.x
pf_mega_location <- pf_mega_location %>%
  select(-title.x, -title.y)

pf_mega_location$text[pf_mega_location$...1 == 2766] <- "The Trans-Pacific trade deal could cost America

write_csv(pf_mega_location, file = "pf_mega_location.csv")
```