PolitiFact Clean Up

Grace Abels

4/19/2022

Version 8 - Removal of PolitiFact anomalous claims

At this stage in the process, we are working with four publisher datasets from which we have sorted out all claims that do not fit our criteria of claims by U.S politician. We have also eliminated Independent and Libertarian claims. Now within each publisher we pivot to focus on textualRating, which denotes the rating that the claim was given by the fact-checker. Some publishers have standardized scales with a ranking system, 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 used to categorize claims.

As stated on 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 they rate statements based on the information known at the time the statement is made."

Below is a breakdown of all the ratings in the PolitiFact dataset's textualRating column.

There are: 2,527 normal claims (those rated with on of the 6 standardized ratings) and 354 anomalous claims (those with a non-standardized text based rating).

Below the breakdown of the 20 most numerous ratings.

Table 1: Textual Ratings

textualRating	n
Half True	555
Mostly True	534
Mostly False	499

textualRating	n
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

When we reviewed the list of anomaly claims, we noticed that they all came from PolitiFact articles, not fact checks. PolitiFact only uses Truth-O-Meter ratings 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 sufficient 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 PolitiFact did not feel comfortable enough to deliver a full Truth-O-Meter rating, we were unwilling to attribute a different level of certainty to the rating. Thus all claims that were not given one the 6 standard ratings in textualRating were removed from the dataset. This resulted in a remaining dataset of 2,527 claims.

Version 9 - Standardized Ratings Only

Table 2: Truth-O-Meter Rating Distribution

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

Version 10 - Condensing claimant names

After this removal we began looking closely at the remaining claimant names. Version 8 had 724 unique claimant names, below are the first 20.

Table 3: Claimant Names by Number of Occurences

claimant	n
Joe Biden	113

claimant	n
Hillary Clinton	104
Bernie Sanders	58
Mike Pence	56
Newt Gingrich	45
Ted Cruz	42
Marco Rubio	39
Ron Johnson	32
Barack Obama	31
Andrew Cuomo	30
Scott Walker	29
Elizabeth Warren	26
Kamala Harris	25
Chris Christie	24
Paul Ryan	24
Rick Scott	24
Tammy Baldwin	24
Nancy Pelosi	23
Tim Kaine	23
Tony Evers	23

At this stage we noticed that some claimants were listed under several separate names despite 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 number 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 claimant namea and then recoded the data accordingly.

This processed combined several names making the list 27 names shorter. 697 unique claimants resulted. The 20 most numerous claimants are listed below.

Table 4: Claimant Names by Number of Occurences, no duplicate names

claimant	n
Joe Biden	113
Hillary Clinton	104
Bernie Sanders	58
Mike Pence	56
Newt Gingrich	45
Ted Cruz	42
Marco Rubio	39
Ron Johnson	32
Andrew Cuomo	31
Barack Obama	31
Scott Walker	29
Rick Scott	28
Elizabeth Warren	26
Kamala Harris	25
Chris Christie	24
Nancy Pelosi	24
Paul Ryan	24
Tammy Baldwin	24

claimant	n
Beto O'Rourke	23
Tim Kaine	23

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, and Marco Rubio's heckler.

21 claims by these claimants 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 (our data said it was spoken by Maxine Waters but the link said it was bloggers) that were not political figures.

For ease we have addressed them at this stage, thus 6 additional claims were removed.

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

Table 5: 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

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

Tagged Claims

To learn more about what politicians were lying about, we subject tagged each claim in the data set with relevant and comprehensive tags to categorize the topics of lies.

We decided to only tag False(ish) claims (Mostly False, False, and Pants on Fire) because out aim is to study what politicians are lying about, not what they are telling the truth about. We acknowledge this choice could limit our inquiry, but since claims were going to be tagged manually we decided to focus on our primary research question. This left 1,085 False(ish) claims to be subject tagged.

Table 6: False(ish) Claims to be Tagged

textualRating	Democratic	Republican
Pants on Fire	31	93
False	160	310

textualRating	Democratic	Republican
Mostly False	218	273

We then had to construct a method for tagging claims. We started, with Frank Baumagrtner's well established subject tags used for the Comparative Agendas Project.

However, since they were designed to tag policy and not politics, there were some missing categories that were politically relevant. To address those missing tags, our research team conducted a series of practice tagging sessions in which 100 randomly selected claims were subject tagged. Through this process we discovered which categories were missing allowed us to establish categories that best encompass the claims found in the dataset. The tagging system and tags themselves were built based on the data we were tagging.

Each tag is defined by a list of topics that fall underneath that subject area. We identified these by buzzwords that may be seen in a claim, or general concepts (more niche than the overall tag) that can help a tagger tag the claims correctly. Despite every effort to be comprehensive, we acknowledge there may be some gaps in our definitions.

Words in italics signify modifications/clarifications to the definition after tagging had begun.

People may disagree with category grouping. We acknowledge the subjective nature of these categories.

Subject	**Or having to do with/associated with**
National and State Macroeconomic Issues	Interest Rates, Inflation, Monetary Policy, Gov. Debt, Economic Regulations, GDP
Economic Well-Being and Domestic Commerce	Banking, Finance, Personal Wealth, Income, Economic Well-Being of Citizens, Poverty, Small Businesses, Stock Market
Labor and Employment	Unions, Labor Force, Unemployment, Jobs, Regulations in this Sphere, Workers Rights, Minimum Wage, Wages, Worker's Benefits
Foreign Trade	Trade Deficits, Tariffs, Imports/Exports, GDP, Trade Deals, Exchange Rates, International Finance
Taxes	Anything to do with Taxes– Tax code, Raising/Lowering Taxes, New Taxes
Government Operations	Gov. Bureaucracy, Legislative Bodies, Gov. Spending, Employees, Appointments, Contracts, Census, Domestic Inter/Intra Governmental Relations, Judiciary, *Impeachment, Executive Orders*
Defense/Military	Military, Military Spending, Armed Forces, Defense Contracts, Weapons, Military Bases, Intelligence, Combat/Wars, Veterans, Military Honors, Eligibility for Service, Cyber Security, National Security
International Affairs and Foreign Aid	Foreign Relations, Alliances, Diplomacy, Human Rights, Development, Embassies Anything International that is not economic
Voting/Elections	Voting Rights, Voter Suppression, Voting policy, Voting Patterns, Elections, Election Fraud, Election Policy, *Gerrymandering, Not Campaigns and Campaign Finance*
Civil Rights Minority Issues and Civil Liberties	Issues of Civil Rights and Liberties– can include issues of Race, Sex, Gender, Religion, Freedom of Speech, Disability, Privacy, Age, Protest
Immigration and Refugee Issues	Illegal/Legal Immigration, Border Policies/Issues, Immigration Policy, Rights of Illegal Immigrants, Visas, Citizenship, Refugees, Child Migrants, Border Patrol

Subject	Or having to do with/associated with
LGBTQ	Sexual Orientation, Gender Presentation, Trans Rights, and Related Policy, Statistics, and Discrimination
Race	Racial Issues, Racial/Ethnic Discrimination, Racial Inequality, Bias, Representation
Religion	Religion, Personal Faith, Religious Freedom, Discrimination based on Religion, Religion in Policy
Women	Women's Rights, Women's Issues, Gender Inequality, Women's Health, Sexism
Health (non-care)	Health outside of Health Insurance, Anything related to human health and well-being, COVID-19, Mortality Rates, Drug Use/Abuse, Mental and Physical Illness, Health Policy
Social Welfare	Social Security, Medicare, Medicaid, Food Stamps. Unemployment Assistance, Disability Assistance, Gov. Programs serving groups in need, Childcare Services, Charities + Volunteer Organizations
Education	Education Policy, Funding, Private/Public School, Early Education, Curriculum, Higher Education, Student Loans, Admissions, Educational Access
Abortion	Pro-Life, Pro-Choice, Fetus Facts, Planned Parenthood, Roe v. Wade, Abortion Clinics, Abortion Policy
Healthcare	Health Insurance, Medicare, Medicaid, Obamacare, Uninsured, Premiums, Cost of Healthcare, Pharmaceuticals, Cost of Medicine, Universal Public Health Care, Cost of Treatment, Wait Times, Death Panels, FDA
Agriculture	Agricultural Subsidies, Animals, Farming, Crops, Agricultural Costs, Food Safety, Regulation, USDA/FDA (food), *Wildlife*
Environment	Environmental Policy, Climate Change, Environmental Disasters, Drinking Water, Waste, Pollution, Recycling, Conservation, EPA, Weather
Energy	Electricity, Coal, Natural Gas, Oil, Nuclear Energy, Renewable Energy, Related Policy
Transportation	Air Travel, Trains, Highways, Infrastructure, Bridges, Public Transport
Community Development and Housing	Community and Urban Development, Housing, Homeless, City Planning, *Population Change*
Technology Science Space	Technology, Tech Policy, Space, Science, Research, Innovation
Media and Communications	Social Media, News Media, Entertainment, Broadcast, Telecommunications, Media Policy, Related Policy
Law and Crime and Policing	Crime, Criminal Justice System, Courts, Police, Violence, Police Brutality, Illegal Drugs, Jails and Prisons, Criminal Justice Reform, Policing Reform, Rights of Felons, Rights of Victims, Mass Incarceration, Crime Rates, *January 6th*
Terrorism	Domestic or International Terrorism, ~~Mass Shootings~~, Interrogation Techniques, Hijacking, Piracy, Policy to Combat Terrorism, Conspiracy
Guns	Gun Rights, Gun Control, Gun Violence, Mass Shootings, 2nd Amendment, Gun Laws
History	Claiming to be "Historic", Citing Historical Events, 20 years ago or older, Precedented/Unprecedented. Generally claims about what has happened in the past.
Record/Candidate Biography/ Campaigns and Personal Behavior	Claims about Personal Behavior, Beliefs, Campaigns, Corruption, Campaign Spending, Scandals. *Only tagged when about a singular person. Not two people or an administration.*

Fear	Claims intended by speaker to elicit fear. Including fear of both immediate or delayed bodily harm, financial/economic harm, harm to property, harm to security, harm to rights, harm to health, harm to democracy. Fear of cultural encroachment. Intended to scare. *Must include inflammatory language, the use of extremes or exaggerations, or current/impending threats of harm to the listener.*
Self/ Personal Record	Claims where the speaker is making a statement about his/herself (including beliefs, record of behavior, political record. Any statement about the claimant themselves.
Opponent/ Opponents Record	Claims where the speaker is making a statement about a political opponent. Can be about what the other person has said, done, not done, beliefs, or character. Opponent is not limited to direct political opponent but all politicians of the opposing party, or the opposing party itself, or those in contradiction with claimant's political position.
Legislation	About a Proposed or Passed Policy, Impact of Policy/Legislation, Nature of Policy, Motivations for Policy. Legislation includes any bills including spending bills and tax bills. Applies when a claim refers to a specific piece of legislation. Not general spending or tax issues.
IMPORTANT NOTES	Claims that are tagged with Self or Opponent are typically also tagged candidate biography, unless the Self/Opponent in question is not an individual but a whole party. Similarly there may be some tagged candidate biography that are not self/Opponent because they are discussing the record of someone within their own party because it is neither their opponent or themselves. Discussion of Taxes and Funding are not always legislation, only is a specific bill/ policy is mentioned. Raising Or Lowering taxes is not unless it is about a tax BILL. Also historic legislation typically does not apply. *Claims about lobbying groups will be tagged according to the issue the group is lobbying for. NRA=Guns etc. If it is about the candidates willingness to respond to lobbyist it is also candidate biography.*

In our tagging of claims, we wanted to be able to describe as many attributes of a claim as possible. First, we allowed a claim to be assigned multiple tags, and we had different types of tags– Macro and Micro.

Micro tags were traditional subject tags that describe the substance of the claim and what political issues or policy issues it is discussing.

Macro tags take a more big picture approach describing the directive of the claim (Opponent/Self/Legislation) and the use of Fear within a claim.

Four different coders were trained on how to tag claims based on the subject tag definitions. Taggers did practice rounds on randomly selected claims and we talked through instances where they disagreed and worked to clarify our definitions.

We were actively seeking consensus between our taggers. Because each claim could receive multiple tags, one different tag would mean a mismatch. High inter-coder reliability would be hard to achieve. Understanding the confounding factors that come with our method of tagging, we feel confident in our system. The tagged data is available for review.

Asa Royal created an online tagging interface for our coders to use. A coder was presented with only the text of a claim, and a link the fact-check if they needed more information. They would then select all the tags they felt applied to that claim. All claims were tagged by at least two coders. If the two coders agreed on the tags, that claim was not reviewed. But, if there was disagreement between the two coders, as was the case for 776 of the claims, lead researcher Grace Abels manually reviewed each claim and the two different sets of tags that had been assigned. Grace Abels then determined the final tags for each claim.

After tagging, the dataset was joined with the larger version 11.

Manual Removal of Duplicate Claims

During tagging, 31 more claims were identified as duplicates, shown below. These were claims with differing urls but identical content. These were not captured earlier because we searched for duplicates with identical urls. Most often these duplicates were fact-checks that had already been published and were included again in an article round-up of fact-checks.

Table 9: 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
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

We prioritized preserving the claims linked to the original and specific fact check over those republished as part of a larger article. To do so, we manually removed the 31 duplicates by their claim identifier number. This resulted in a remaining 2469 claims.

Location of Lie

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. There was one claim that was erroneously untagged with location information, this was fixed. Then, we exported the final dataset into a viewable and downloadable csv, called `pf_mega_location.csv`.

Further analysis of this data can be found in the Location of Lie Analysis, available in both accessible RMD and PDF form.