LIAR, LIAR, PANTS ON FIRE:

Disinformation in American Media (2007-2016)

By: Gaby Gerecht, Gloria Gerhardt, Senandung Luluk, Hugo Hsieh

FAKE NEWS

Automatic fake news detection is a challenging problem in deception detection, and it has tremendous real world political and social impacts

“Right before the 2016 election, the top 20 fake news stories that were circulating on social media received more engagement — so that’s liking, sharing, commenting — than the top 20 factual news stories that were on social media” - Chrysalis Wright, UN’s Communications Coordination Committee Member

FAKE NEWS

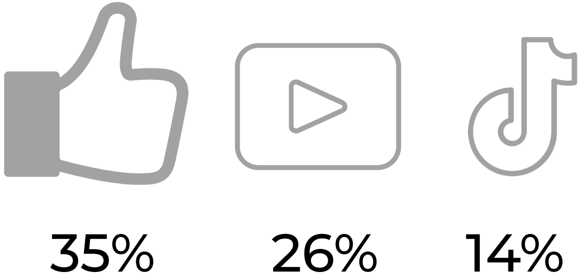
Key facts about American usage of social media

HOW

86% of U.S. adults say they often or sometimes get **news from a digital source**

WHERE

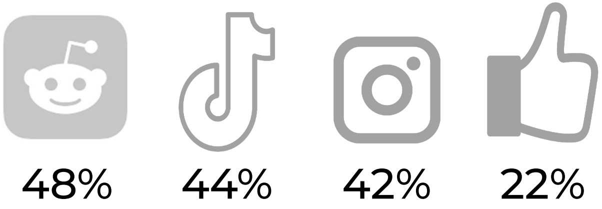
30% of Americans get news regularly or sometimes from:



WHO

Ages 18-29

69% get their news at least sometimes from social media – this declines with age



FAKE NEWS

Disinformation poses a threat to democracy in 2024

50%

Over half of Americans claim to **regularly see fake** news on social media

Research suggests disinfo has little direct effect on voting choices, **but spread by political elites**, it can impact how people decide on key issues.

1/4

One fourth of Americans **don't trust the news** on social media

57% of Republican voters

Believed that the 2020 election was stolen

A lack of voter trust in elections can lead to **violence**

64%

Of election officials reported in 2022 that the spread of false information has made their **jobs more dangerous**

FAKE NEWS

Disinformation policy issue through LIAR

VS

- **WHO** spreads fake news in blue vs red states
- **HOW** much fake news is spread in blue vs red states

- **FALSE RATES** in 2016 vs 2024 swing states
- **ELECTION COMPETITIVENESS** and FALSE RATES

LIAR DATASET

The LIAR dataset is a resource for fake news detection from POLITIFACT.COM.

It is significantly larger than other public datasets in the field. Each statement has been meticulously evaluated for truthfulness by POLITIFACT editors.

- 12.8K short statements
- Manually labeled over a decade
- Six labels:
  Pants-fire, FALSE, Mostly-false,
  Half-true, Mostly-true, TRUE

LIAR STATS

Dataset Statistics

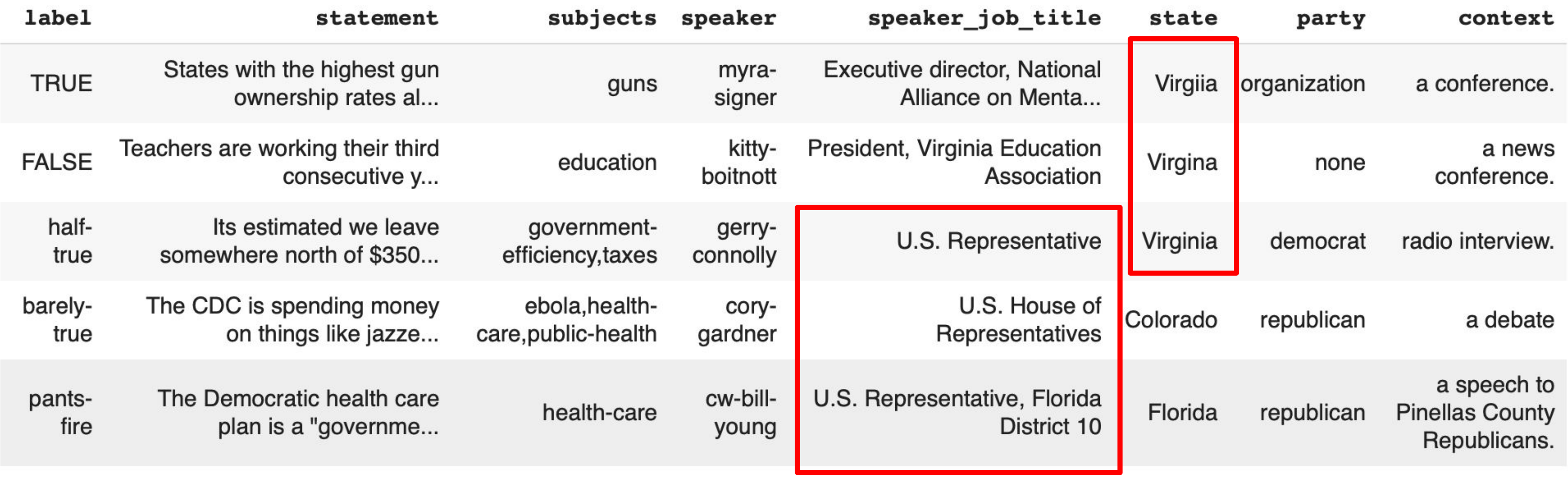
Training set size	10,240
Validation set size	1,284
Testing set size	1,267
Avg. statement length (tokens)	17.9

Top-3 Speaker Affiliations

Democrats	4,137
Republicans	5,665
None (e.g., FB posts)	2,181

True	2,053
Mostly true	2,454
Half true	2,627
Barely true	2,103
False	2,507
Pants-on-fire	1,047

LIAR PREVIEW



label	statement	subjects	speaker	speaker_job_title	state	party	context
TRUE	States with the highest gun ownership rates al...	guns	myra-signer	Executive director, National Alliance on Menta...	Virgiia	organization	a conference.
FALSE	Teachers are working their third consecutive y...	education	kitty-boitnott	President, Virginia Education Association	Virgina	none	a news conference.
half-true	Its estimated we leave somewhere north of $350...	government-efficiency,taxes	gerry-connolly	U.S. Representative	Virginia	democrat	radio interview.
barely-true	The CDC is spending money on things like jazze...	ebola,health-care,public-health	cory-gardner	U.S. House of Representatives	Colorado	republican	a debate
pants-fire	The Democratic health care plan is a "governme...	health-care	cw-bill-young	U.S. Representative, Florida District 10	Florida	republican	a speech to Pinellas County Republicans.

We will focus on exploratory data analysis and visualization.

LIAR LIMITATIONS

Key limitations of the dataset

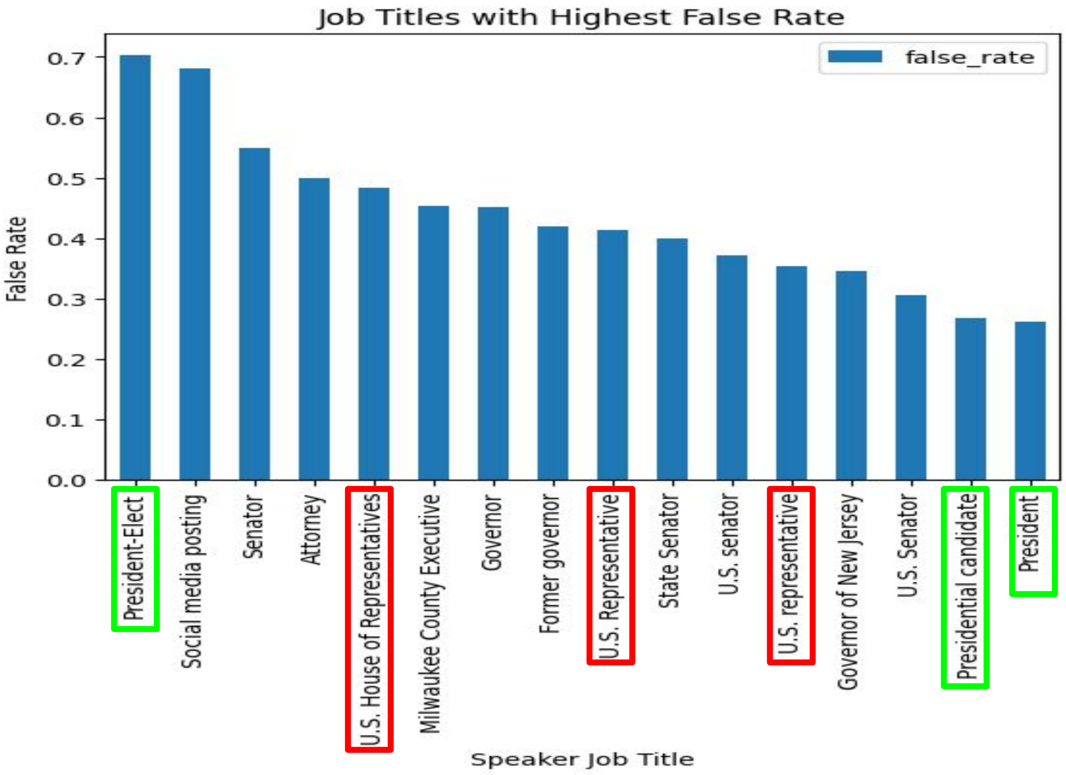
1. **Outdated**
   The data set only has statements made from 2007 - 2016.

2. **Statements are not dated**
   The dataset does not include dates for when each statement was made.

3. **Highly unorganized and messy data**
   N/A, Lack of categorization, Misspelling

LIAR LIMITATIONS

Before cleaning

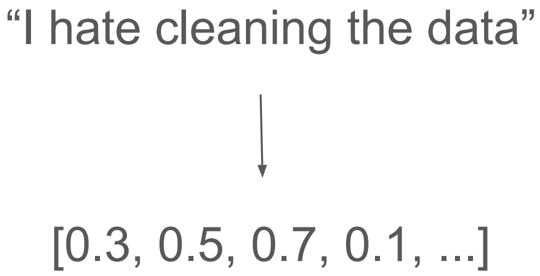


METHODOLOGY

How can we organize the data (if not by hand)

Step 1

Vectorize sentences

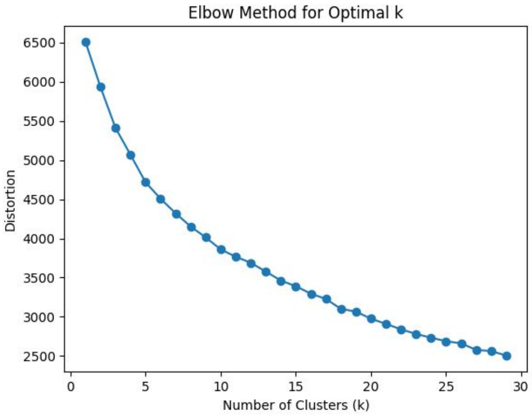


“I hate cleaning the data”

[0.3, 0.5, 0.7, 0.1, ...]

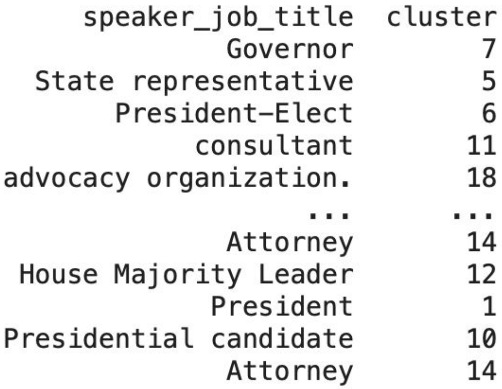
Step 2

K-means Clustering



Step 3

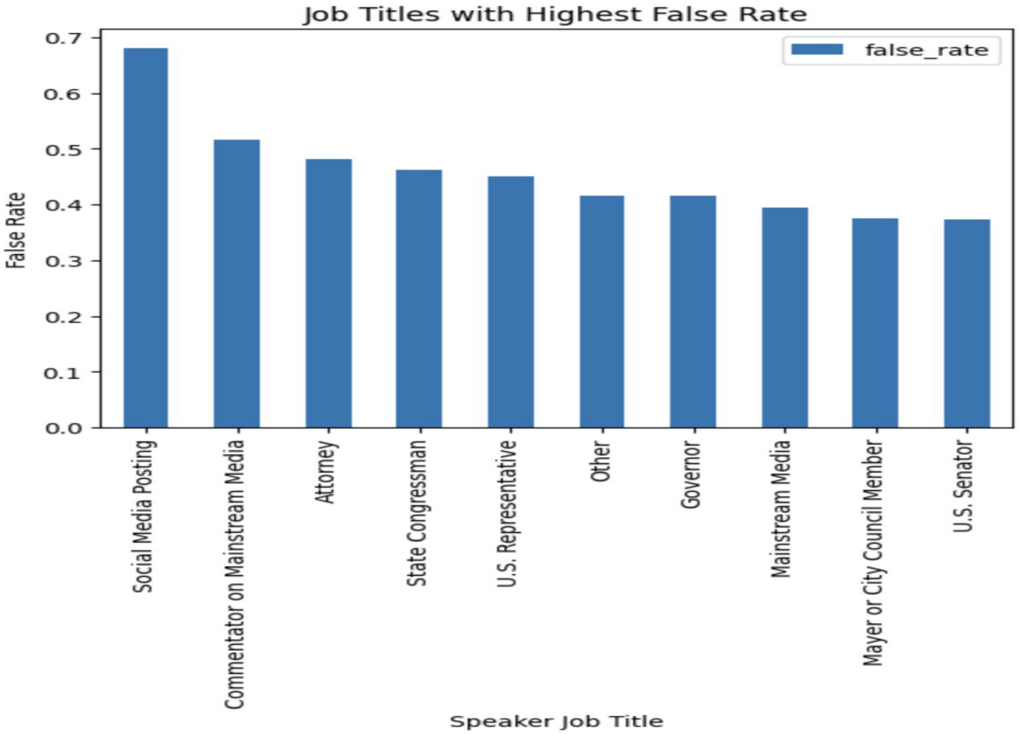
Name the clusters



```
           speaker_job_title  cluster
                    Governor        7
        State representative        5
              President-Elect        6
                  consultant       11
    advocacy organization.        18
                         ...      ...
                    Attorney       14
         House Majority Leader     12
                   President        1
       Presidential candidate      10
                    Attorney       14
```

LIAR LIMITATIONS

After cleaning



METHODOLOGY

How we looked at the data

1. **Concatenate Dataset**

   Concat training, validation, and test set

2. **Adjust Label**

   [‘Pants-fire’, ‘FALSE’, ‘barely-true’]=false; [‘half-true’, ‘mostly-true’, ‘TRUE’]=not_false

3. **Set Threshold**

   Don’t want 1 / 1 = 100%

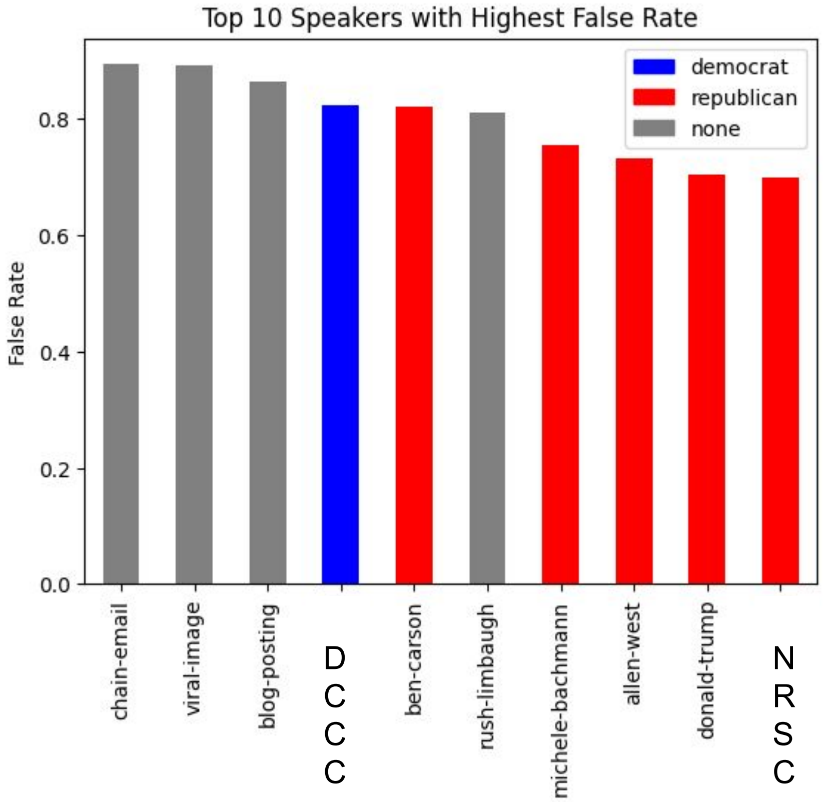
4. **Calculate False Rate**

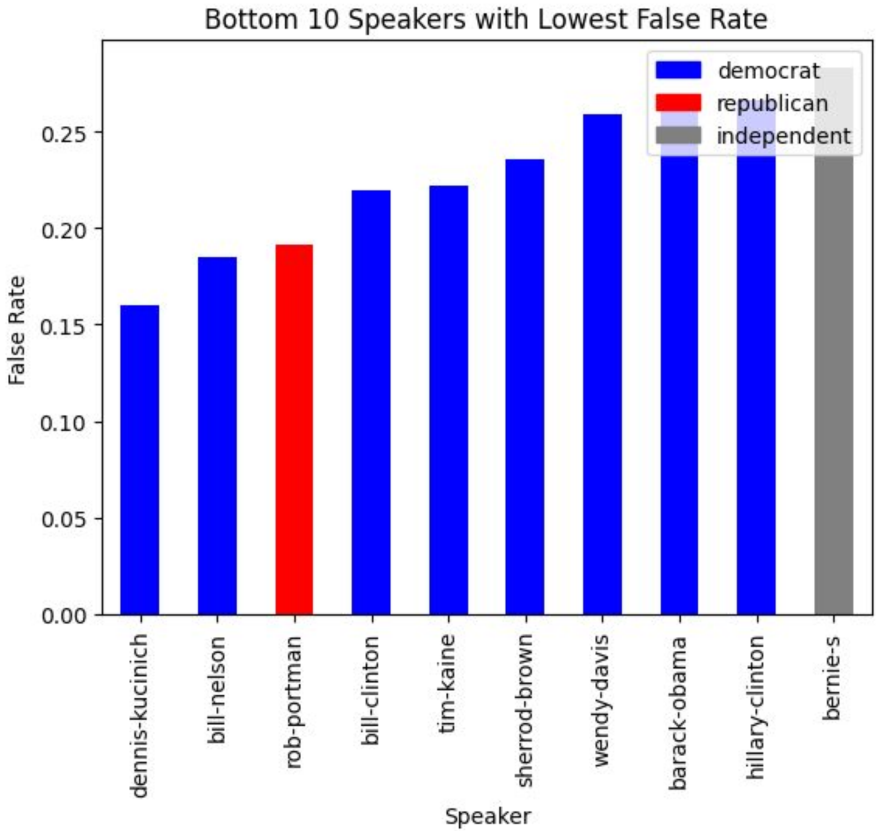
   Don’t want counts; False rate = # of false / # of all

5. **Sort and plot**

   Sort by false rate and make bar plot

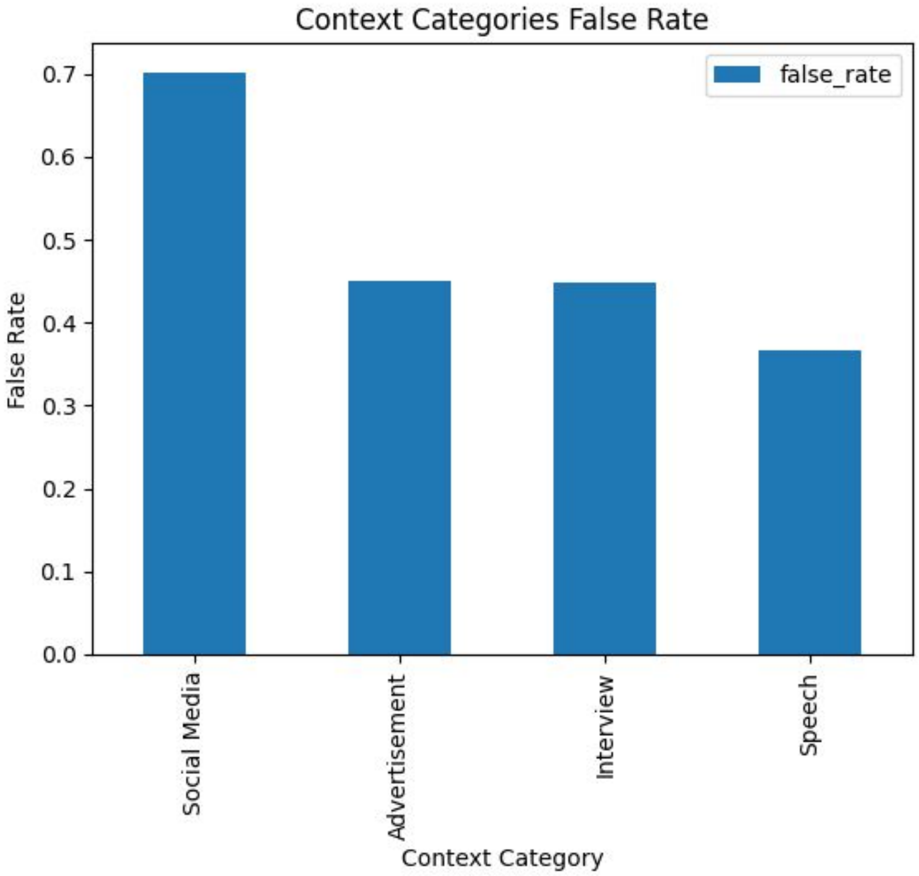
GENERAL FINDINGS



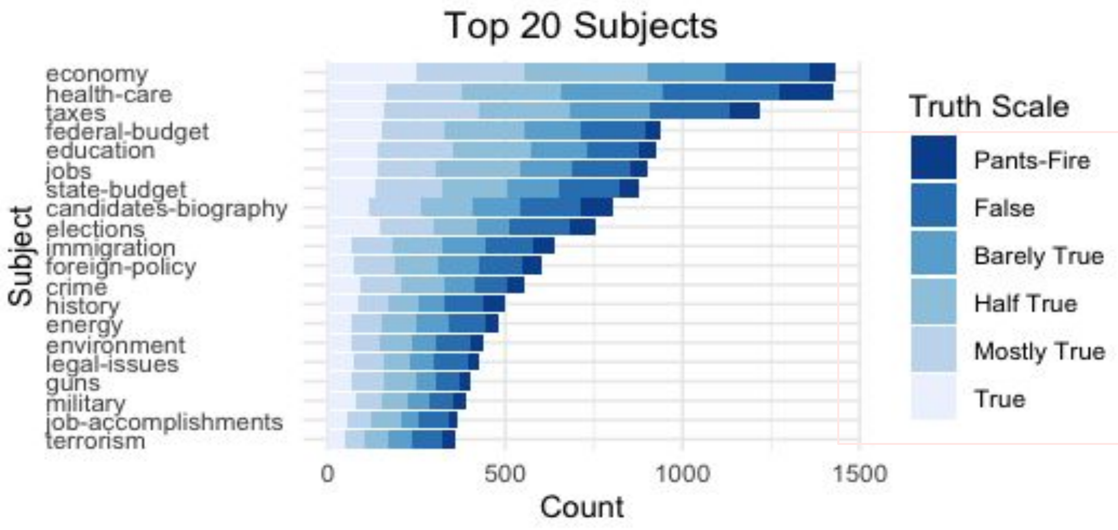


DCCC=Democratic Congressional Campaign Committee
NRSC=National Republican Senatorial Committee

GENERAL FINDINGS



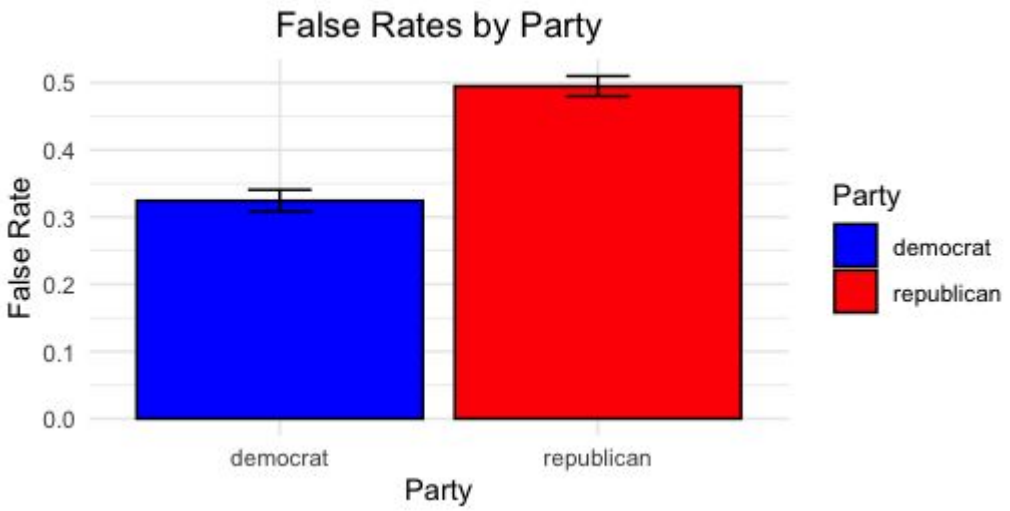
GENERAL FINDINGS



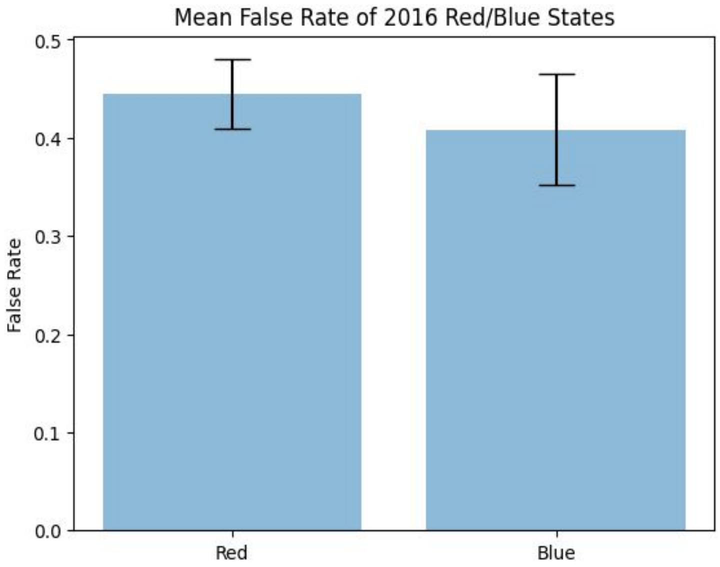
- A majority of statements in the dataset are counted as half-true
- Least number of statements are counted as pants-on-fire
- Top subject in the dataset is the economy

KEY FINDINGS

Republican vs democrat false rate



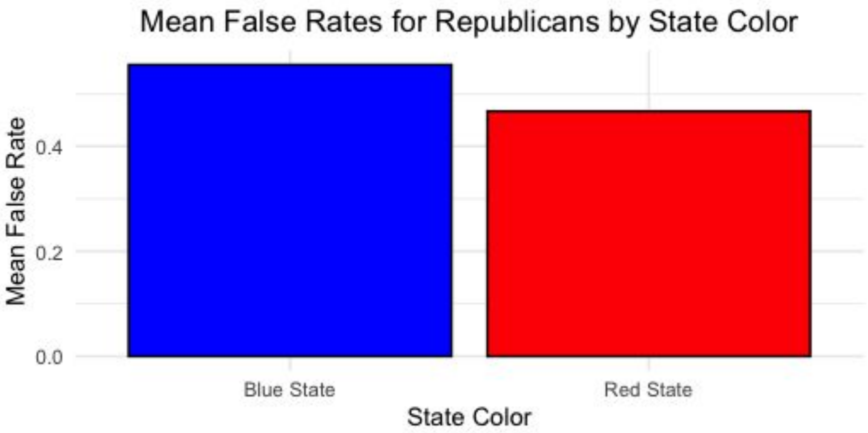
p-value: < 2.2e-16

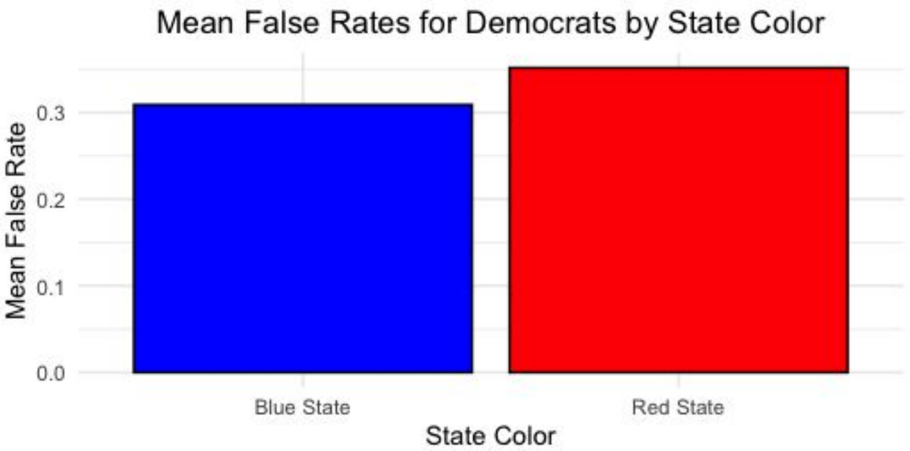


P-value: 0.2761

KEY FINDINGS

Fake news spread in red vs blue states



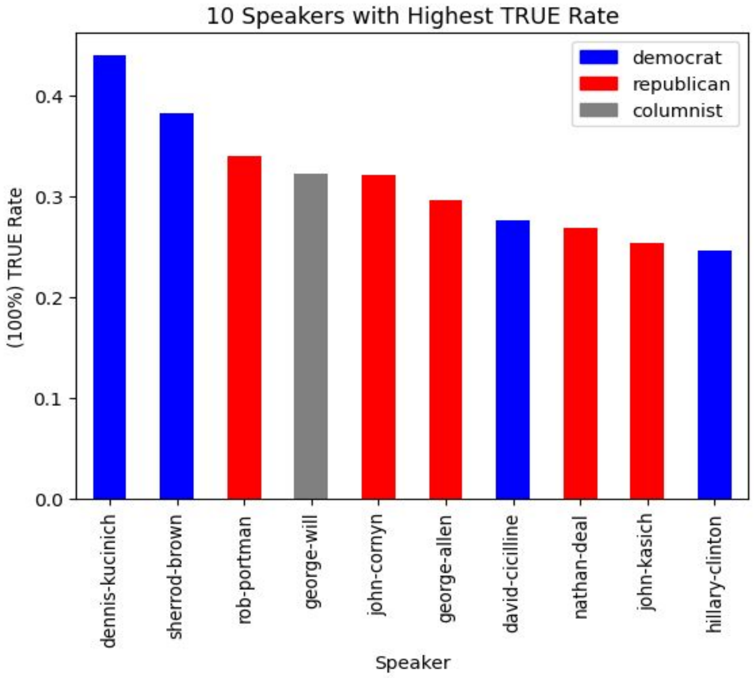


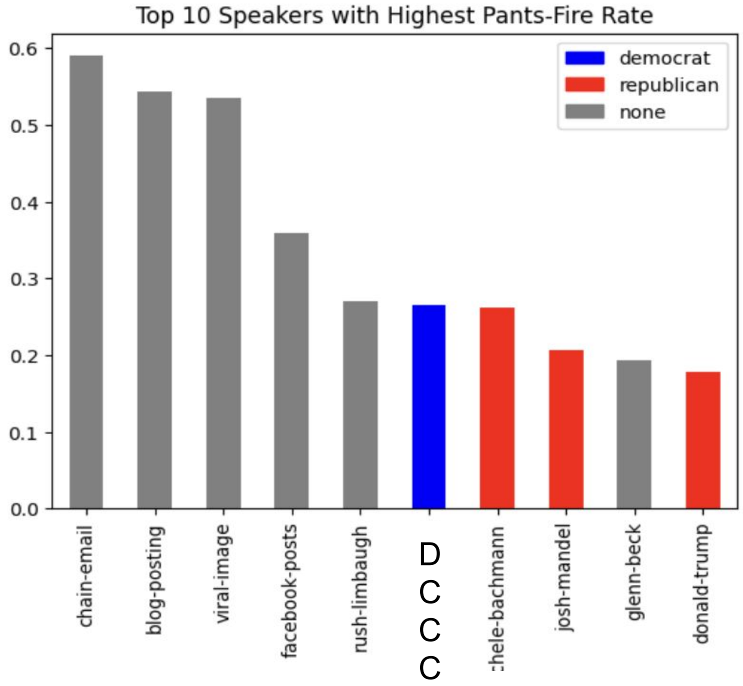
p-value = 7.585e-08

p-value = 0.0138

KEY FINDINGS

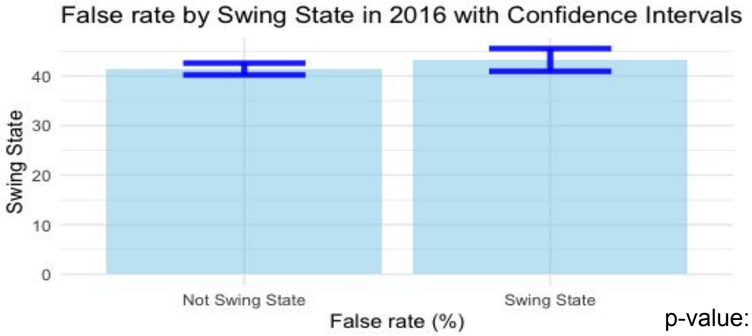
Top politicians true vs. pants-on-fire statements



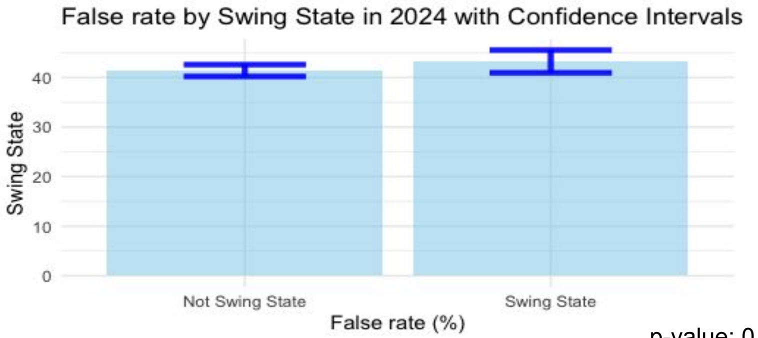


KEY FINDINGS

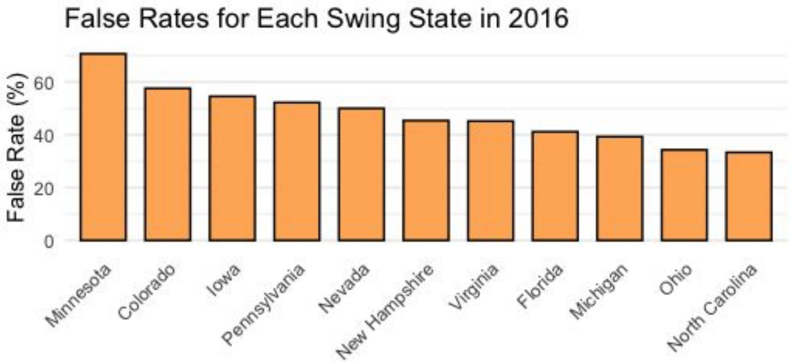
False rates in 2016 and 2024 swing states

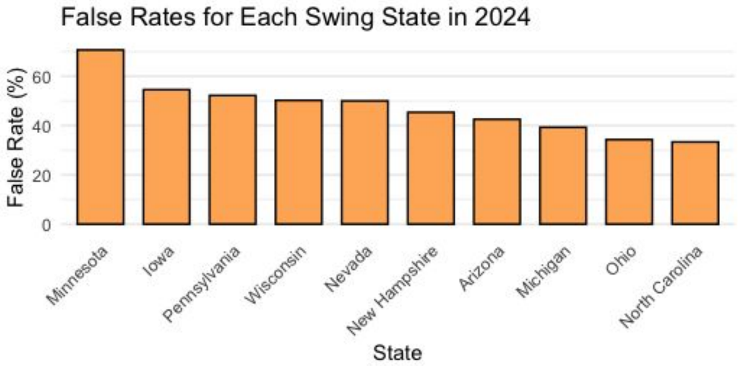


p-value: 0.5341



p-value: 0.1618

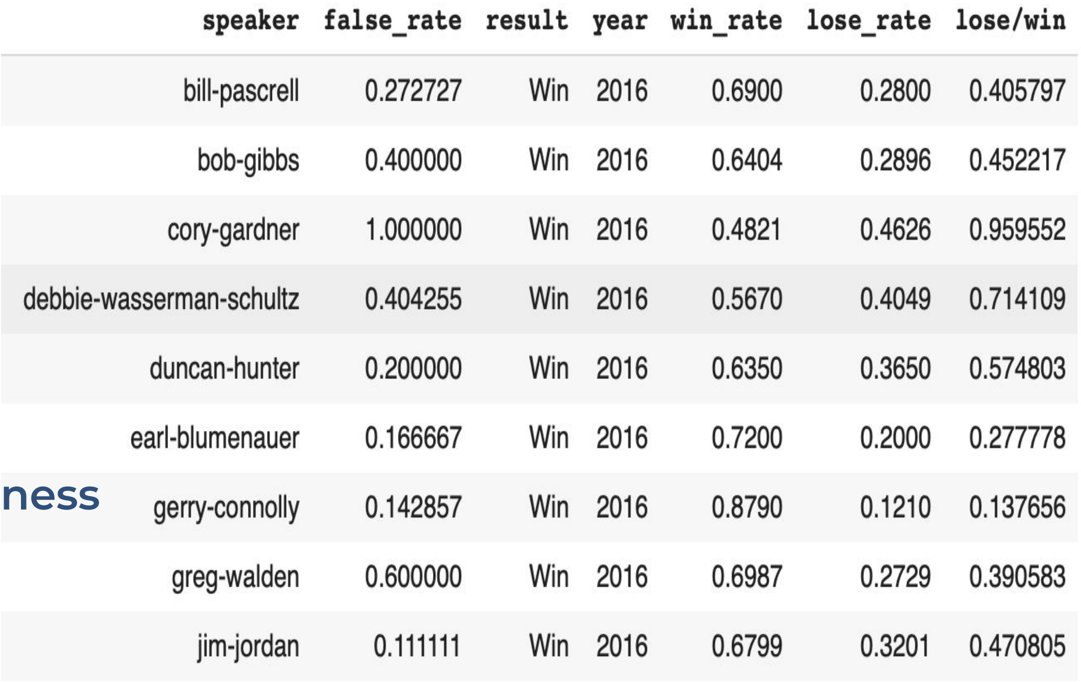




METHODOLOGY

False rate trends based on election competitiveness

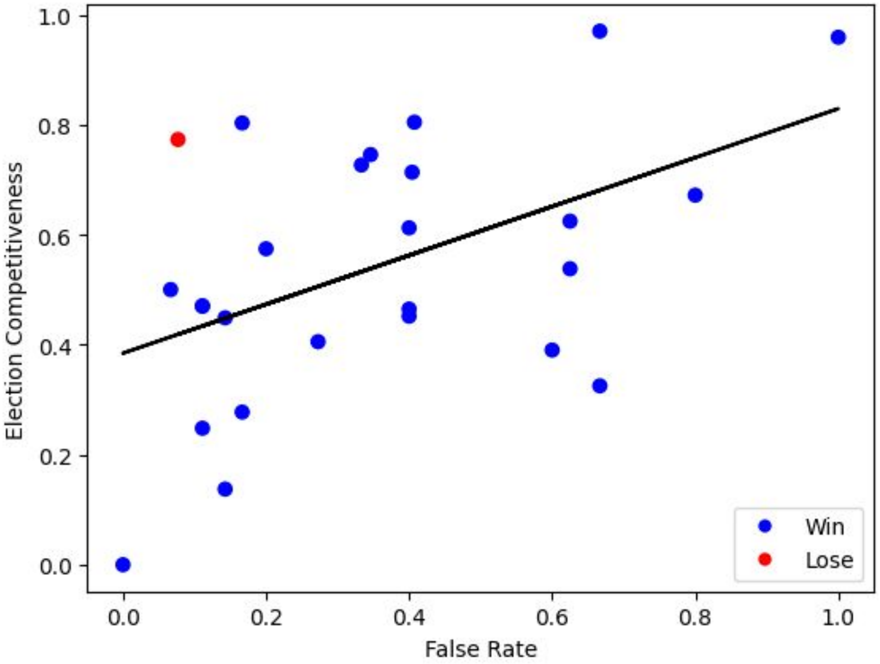
1. Retrieve representatives and senators' false rate from LIAR
2. Run for 2016 election?
3. Percentage of votes from Wikipedia
4. Calculate lose/win ratio to evaluate election competitiveness
5. Regress lose/win ratio on false rate

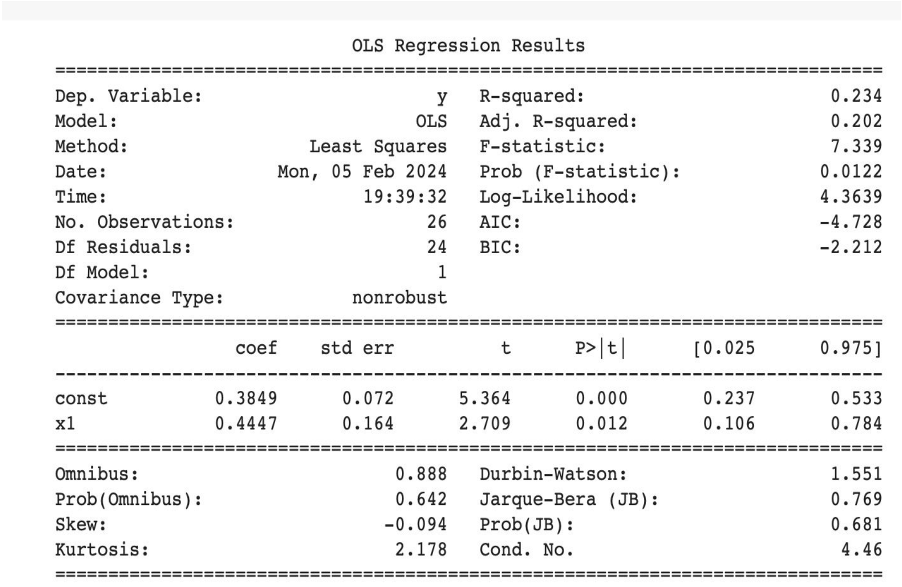


speaker	false_rate	result	year	win_rate	lose_rate	lose/win
bill-pascrell	0.272727	Win	2016	0.6900	0.2800	0.405797
bob-gibbs	0.400000	Win	2016	0.6404	0.2896	0.452217
cory-gardner	1.000000	Win	2016	0.4821	0.4626	0.959552
debbie-wasserman-schultz	0.404255	Win	2016	0.5670	0.4049	0.714109
duncan-hunter	0.200000	Win	2016	0.6350	0.3650	0.574803
earl-blumenauer	0.166667	Win	2016	0.7200	0.2000	0.277778
gerry-connolly	0.142857	Win	2016	0.8790	0.1210	0.137656
greg-walden	0.600000	Win	2016	0.6987	0.2729	0.390583
jim-jordan	0.111111	Win	2016	0.6799	0.3201	0.470805

KEY FINDINGS

False rate trends based on election competitiveness





```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.234
Model:                            OLS   Adj. R-squared:                  0.202
Method:                 Least Squares   F-statistic:                     7.339
Date:                Mon, 05 Feb 2024   Prob (F-statistic):             0.0122
Time:                        19:39:32   Log-Likelihood:                 4.3639
No. Observations:                  26   AIC:                            -4.728
Df Residuals:                      24   BIC:                            -2.212
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3849      0.072      5.364      0.000       0.237       0.533
x1             0.4447      0.164      2.709      0.012       0.106       0.784
==============================================================================
Omnibus:                        0.888   Durbin-Watson:                   1.551
Prob(Omnibus):                  0.642   Jarque-Bera (JB):                0.769
Skew:                          -0.094   Prob(JB):                        0.681
Kurtosis:                       2.178   Cond. No.                         4.46
==============================================================================
```

KEY LESSONS

The LIAR dataset offers several key lessons

VS

1. **Most** fake news is spread on **social media**. The **least** amount of fake news is spread on **traditional media** (tv, print).
   - Trends show that Americans are increasing the amount of news they get from social media while consumption of traditional media is decreasing

2. From 2007 - 2016, republican politicians have spread **more** fake news than democrats
   - HOWEVER, the same amount of fake news was spread in red and blue states

3. Democrats spread more fake news in red states WHEREAS republicans spread more fake news in blue states

4. The more the candidate is in a **competitive election**, the more they are likely to spread news with **a higher** false rate

KEY LESSONS

The LIAR dataset offers several key lessons

1. **Same amount** of fake news spread in swing and non-swing states

2. In the 2016 swing states, Minnesota and Colorado had the highest amounts of fake news in the dataset

3. In the 2024 swing states, **Minnesota and Iowa** had the highest amounts of fake news in the dataset

RECOMMENDATIONS

Developing and applying lessons

Congressmen

Partisan animosity drives news sharing
Greater bipartisan support for anti-disinformation policies, like the Local Journalism Sustainability Act

Increase support to fact-checking services
Support and expand independent fact-checking organizations that can provide real-time verification of claims made by politicians, public figures, and news media

Local Politicians

Digital & civic literacy is imperative
Provide constituents with the skills to access, analyze, and act on digital information based on new standards for digital and civic literacy

Swing state election workers need to be prepared
Minnesota Secretary of State Steve Simon’s office is spearheading, #TrustedInfo2024, an online public education effort to promote election officials as a trusted source of election information in 2024

RECOMMENDATIONS

Developing and applying lessons

Media

Mainstream media is still the most truthful form of media.

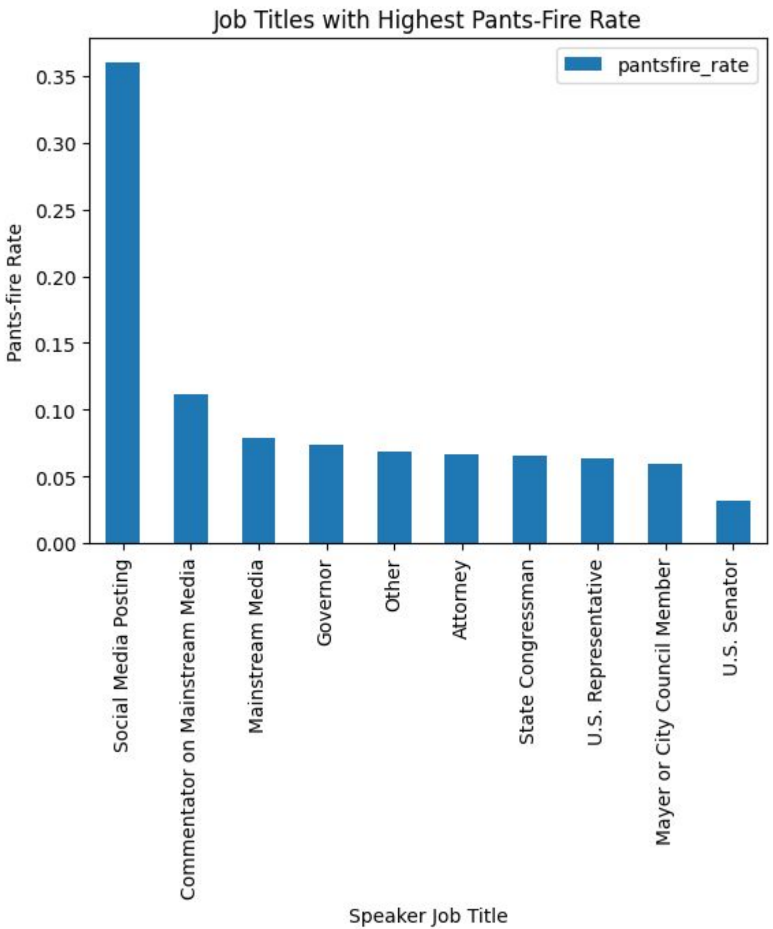
The media should develop industry wide standards on how to disclose the ways they collect, report, and disseminate the news.

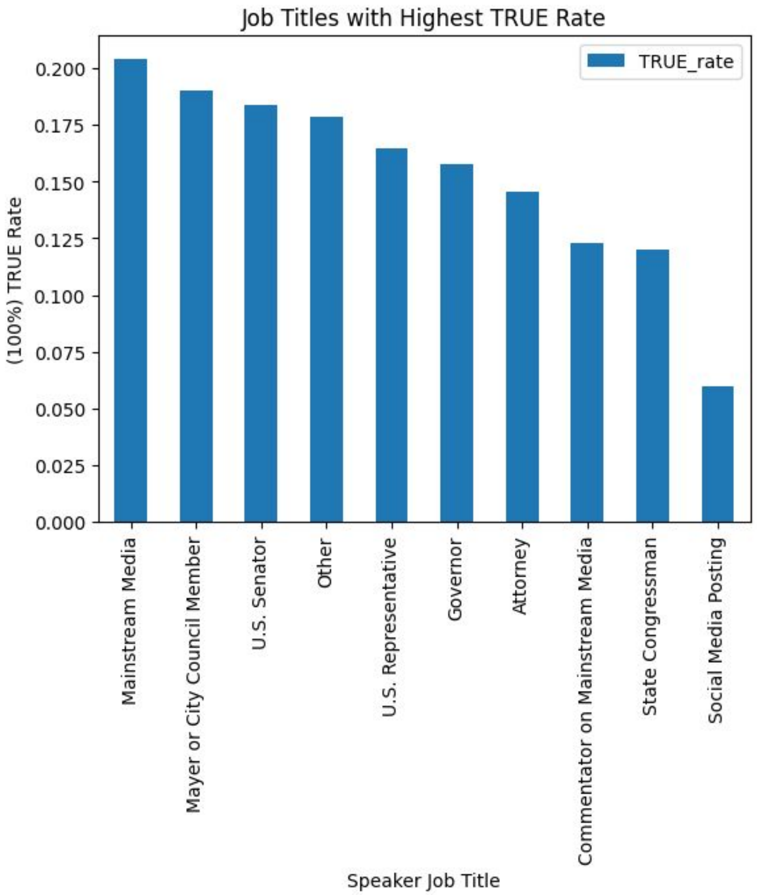
Media and technology companies must be able to determine and then address disinformation while exposing their audiences to diverse viewpoints, particularly in states vulnerable to fake news.

In 2021, Americans were 17 points more likely to trust reporting by local news

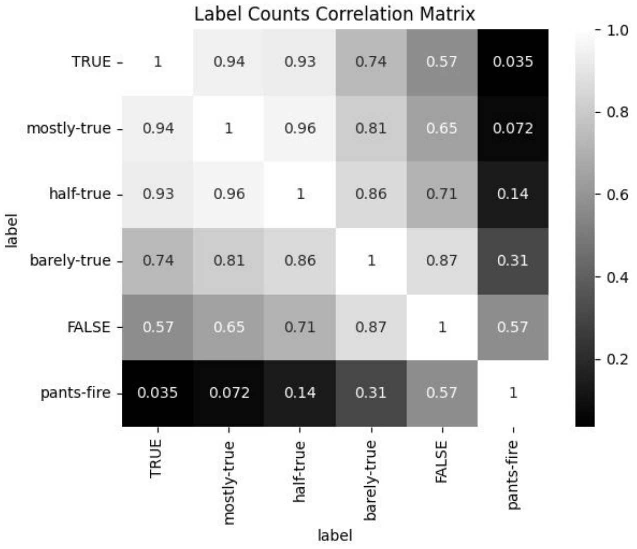
Greater investments in local news agencies, particularly in key swing states where disinformation is highest. Investments can be in digital education for new agency staff, advertising for local news etc...

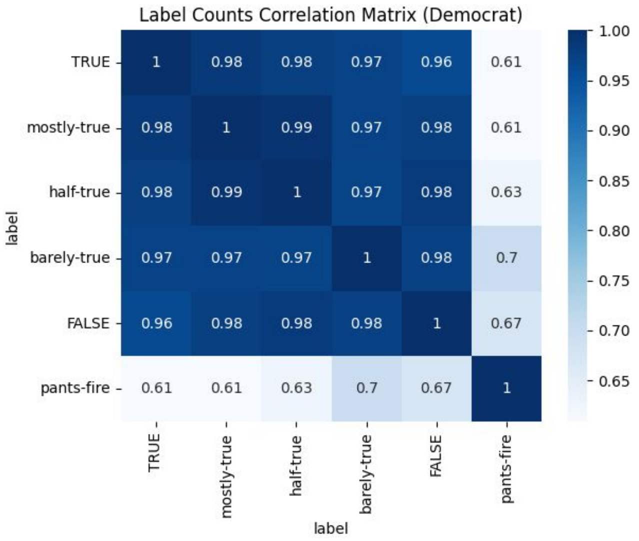
ANNEX

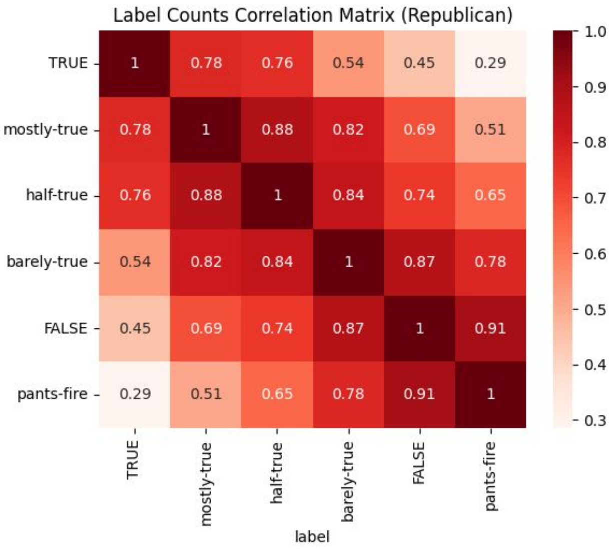

Job Titles with Highest Pants-Fire Rate
pantsfire_rate
Pants-fire Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Social Media Posting
Commentator on Mainstream Media
Mainstream Media
Governor
Other
Attorney
State Congressman
U.S. Representative
Mayer or City Council Member
U.S. Senator
Speaker Job Title


Job Titles with Highest TRUE Rate
TRUE_rate
(100%) TRUE Rate
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Mainstream Media
Mayer or City Council Member
U.S. Senator
Other
U.S. Representative
Governor
Attorney
Commentator on Mainstream Media
State Congressman
Social Media Posting
Speaker Job Title

ANNEX


Label Counts Correlation Matrix
TRUE
mostly-true
half-true
barely-true
FALSE
pants-fire
label
1 0.94 0.93 0.74 0.57 0.035
0.94 1 0.96 0.81 0.65 0.072
0.93 0.96 1 0.86 0.71 0.14
0.74 0.81 0.86 1 0.87 0.31
0.57 0.65 0.71 0.87 1 0.57
0.035 0.072 0.14 0.31 0.57 1
1.0
0.8
0.6
0.4
0.2


Label Counts Correlation Matrix (Democrat)
TRUE
mostly-true
half-true
barely-true
FALSE
pants-fire
label
1 0.98 0.98 0.97 0.96 0.61
0.98 1 0.99 0.97 0.98 0.61
0.98 0.99 1 0.97 0.98 0.63
0.97 0.97 0.97 1 0.98 0.7
0.96 0.98 0.98 0.98 1 0.67
0.61 0.61 0.63 0.7 0.67 1
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65


Label Counts Correlation Matrix (Republican)
TRUE
mostly-true
half-true
barely-true
FALSE
pants-fire
label
1 0.78 0.76 0.54 0.45 0.29
0.78 1 0.88 0.82 0.69 0.51
0.76 0.88 1 0.84 0.74 0.65
0.54 0.82 0.84 1 0.87 0.78
0.45 0.69 0.74 0.87 1 0.91
0.29 0.51 0.65 0.78 0.91 1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

Statement of contribution:

Gaby & Gloria were responsible for the policy part of the presentation while Luluk & Hugo were in charge of processing the data. Gloria mostly conducted the background research on Americans’ use of social media and impacts of misinformation on elections (with some assistance from Gaby.) In the meantime, Luluk and Hugo cleaned the data and ran many different analyses to get an idea of which information we can generate with the dataset. After that, the four of us together looked at the results from these analyses to decide which ones we wanted to include in our presentation to create a coherent story. Afterwards, Gaby developed the policy recommendations (with some assistance from Gloria) and designed the layout of the powerpoint presentation.

Luluk and Hugo collaborated on data cleaning (removing NA, renaming data points, creating categorization, merging data) to perform analysis and visualization. While Luluk focused on analyzing distribution of false rate across different subjects, parties, and states (red/blue, swing/non-swing), Hugo put more effort on investigating distribution of false/TRUE/pants-fire rate across different speakers, contexts, and job-titles. In addition, Hugo contributed on collecting data and conducting statistical inference to investigate association between false rate and election competitiveness of the 2016 election.

Use of AI:

Gloria: I used deepL to double check spelling/ grammar of some sentences

Luluk: ChatGPT used to generate code for plotting and checking/correction of multiple code (i.e renaming states name, creating binary red/blue state, swing/non swing state, etc)

Hugo: I used ChatGPT to help find misspellings in the dataset, generate some codes for data manipulation and visualization, and debug the code I wrote.

Gaby: I used a PowerPoint GPT to help visualize how the slides could look like. I also used it to help create the powerpoint slide titles