Modeling and Measuring Expressed (Dis)belief in (Mis)information

Shan Jiang, Miriam Metzger, Andrew Flanagin, Christo Wilson

Northeastern University

UC SANTA BARBARA

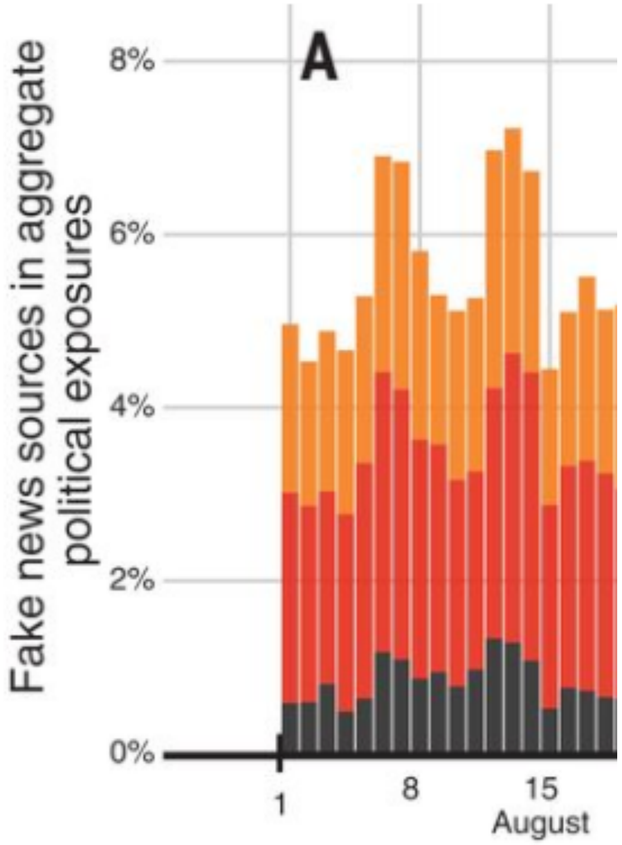
Background: the misinformation problem

(Mis)information

Background: the misinformation problem

2016 US Presidential Election:

- 6% of all news consumptions are “fake news” [1].

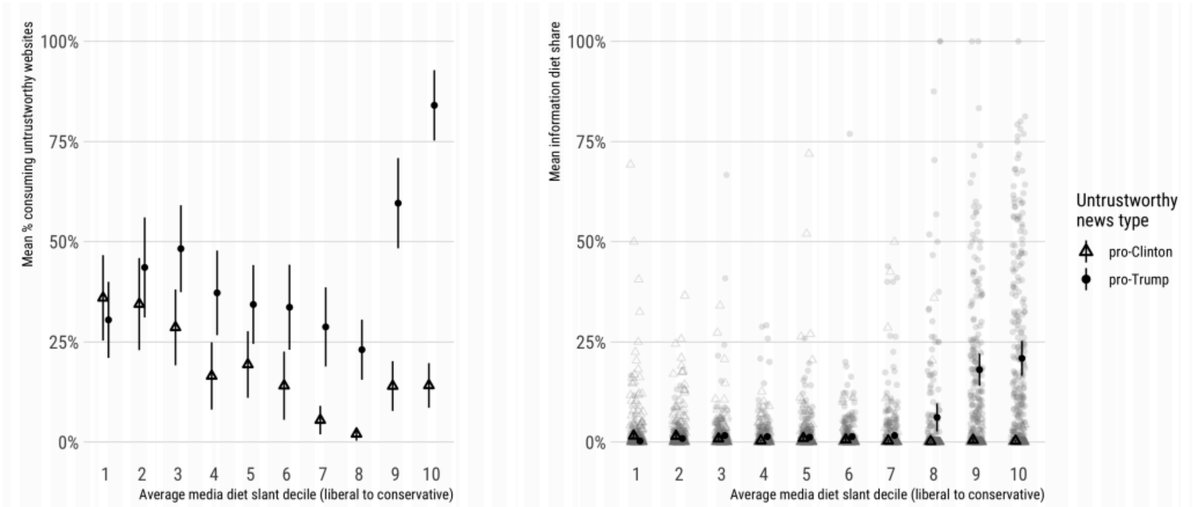


[1] Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on twitter during the 2016 us presidential election. *Science.*

Background: the misinformation problem

2016 US Presidential Election:

- 6% of all news consumptions are “fake news” [1].
- 44% of Americans visited at least one untrustworthy website [2].



[2] Guess, A.; Nyhan, B.; and Reifler, J. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *ERC*.

Background: the misinformation problem

2016 US Presidential Election:

- 6% of all news consumptions are “fake news” [1].
- 44% of Americans visited at least one untrustworthy website [2].

COVID-19 Pandemic:

- Chinese/US biological weapon.
- 5G tower emission.
- Drinking bleach.

Background: believe or disbelieve?

2016 US Presidential Election:

- 6% of all news consumptions are “fake news” [1].
- 44% of Americans visited at least one untrustworthy website [2].

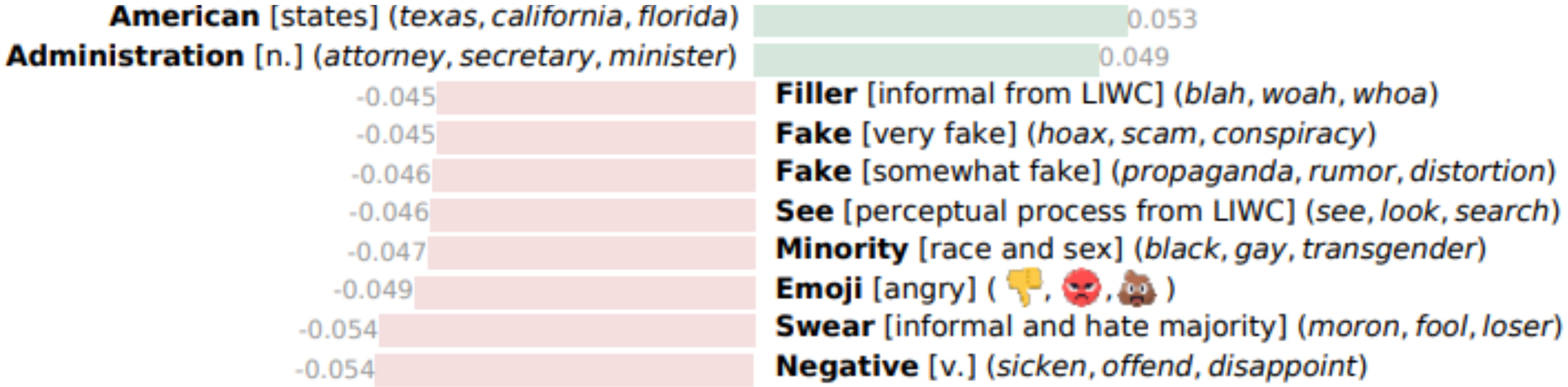
Believe or disbelieve?

COVID-19 Pandemic:

- Chinese/US biological weapon.
- 5G tower emission.
- Drinking bleach.

Background: believe or disbelieve?

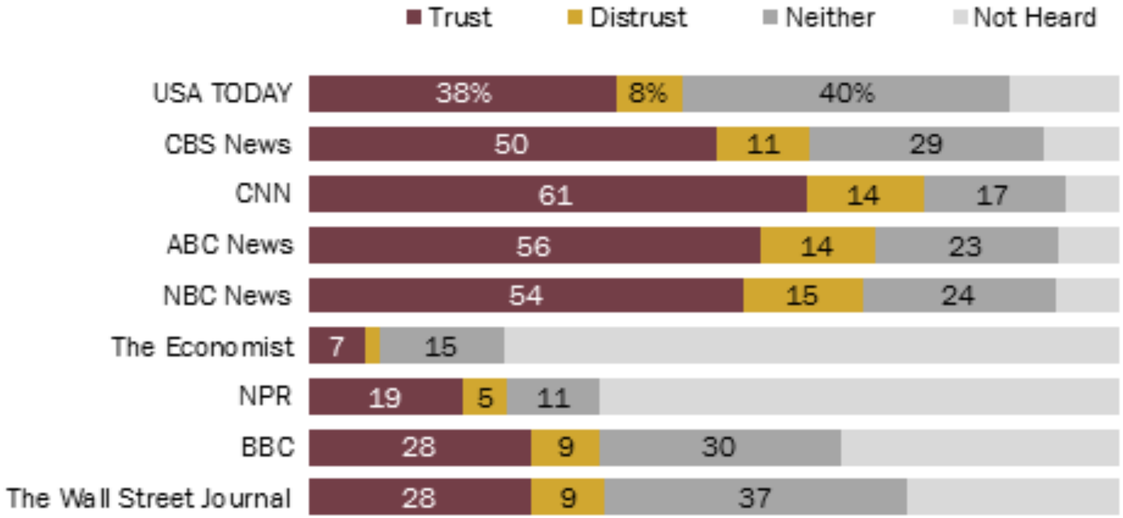
- People use more words indicating their awareness of misinformation in response to false claims than truthful ones [3].



[3] Jiang, S., and Wilson, C. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *PACMHCI (CSCW)*.

Background: believe or disbelieve?

- People use more words indicating their awareness of misinformation in response to false claims than truthful ones [3].
- People believe in some news outlets more than others [4].



[4] Anderson, J., and Rainie, L. 2017. The future of truth and misinformation online. *Pew Research Center*.

Background: research questions

Computational methods to model and measure (dis)belief.

Background: research questions

Computational methods to model and measure (dis)belief.

- Overall prevalence of (dis)belief.
- Effect of time.
- Effect of fact-checking.
- ...

Background: social media comments

Claim

We have now Tested more than 5 Million People. That is more than any other country in the World, and even more than all major countries combined!

Background: social media comments

Claim

We have now Tested more than 5 Million People. That is more than any other country in the World, and even more than all major countries combined!

Belief

That is a vast amount of testing in such a short period of time!

Thank you Mr. President for working so hard for all Americans.

Background: social media comments

Claim

We have now Tested more than 5 Million People. That is more than any other country in the World, and even more than all major countries combined!

Belief

That is a vast amount of testing in such a short period of time!

Thank you Mr. President for working so hard for all Americans.

Disbelief

This is fewer tests than just Russia, Germany, and Italy have done.

…it's plain wrong. Something in the order of 30million global tests done.

Background: steps

- Build a *small* and *labeled* dataset.
- Conduct experiments to model the dataset.
- Apply models on a *large* and *unlabeled* dataset.

Data: comments for claims

- Use fact-checks from PolitiFact between Jan to Jun, 2019, whose claims were originated from Twitter.

Data: comments for claims

- Use fact-checks from PolitiFact between Jan to Jun, 2019, whose claims were originated from Twitter.
- Query an archived 1% sample of tweets and find all comments to the claim tweets.
- 18 claims, 6,809 comments.

Data: annotating (dis)belief

- Each annotator is asked to provide binary labels for each comment: disbelief (0/1) and belief (0/1).

Data: annotating (dis)belief

- Each annotator is asked to provide binary labels for each comment: disbelief (0/1) and belief (0/1).

…it's plain wrong. Something in the order of 30million global tests done.	disbelief: 1 belief: 0
I just want a President who doesn't tell the Americans to drink bleach.	disbelief: 0 belief: 0

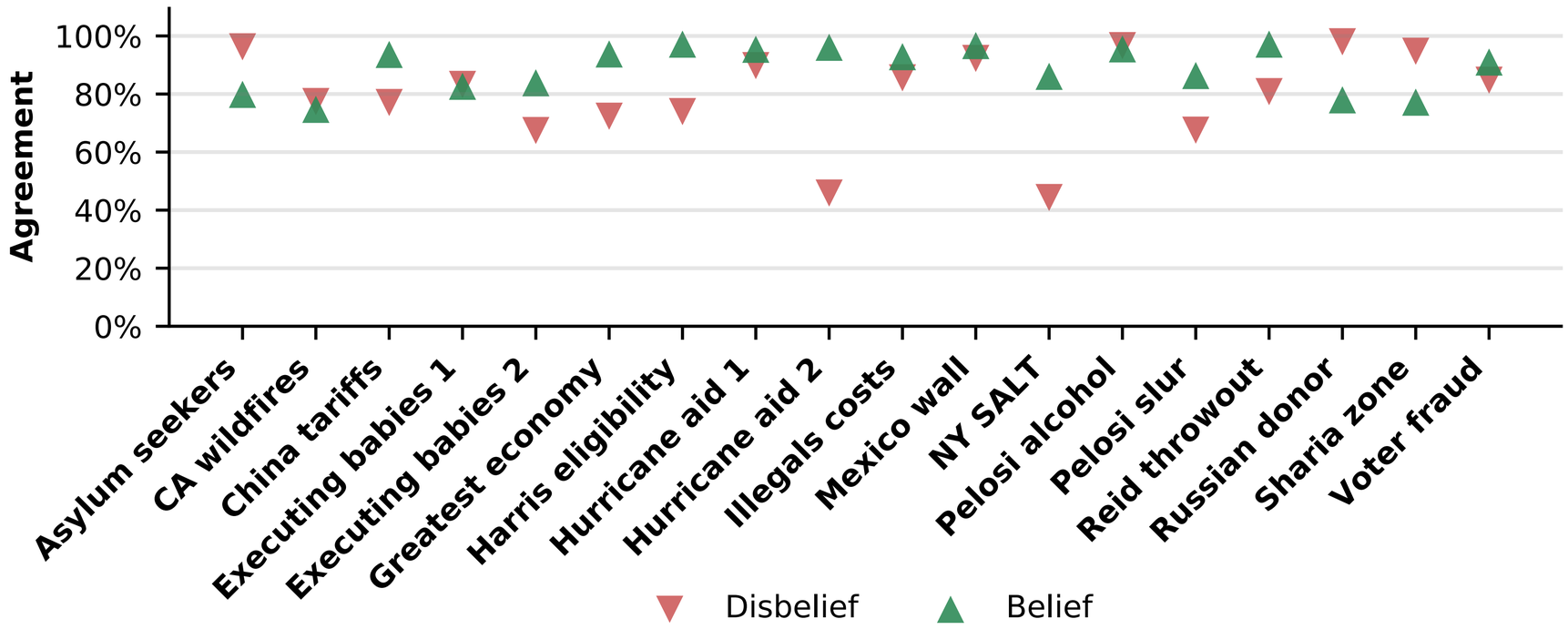
Data: annotating (dis)belief

- Each annotator is asked to provide binary labels for each comment: disbelief (0/1) and belief (0/1).

- 2 annotators for each claim.
- 3rd annotator to break ties.

Data: inter-annotator agreement

- Of 36 claims * labels: 24 above 80%, 32 above 70%.

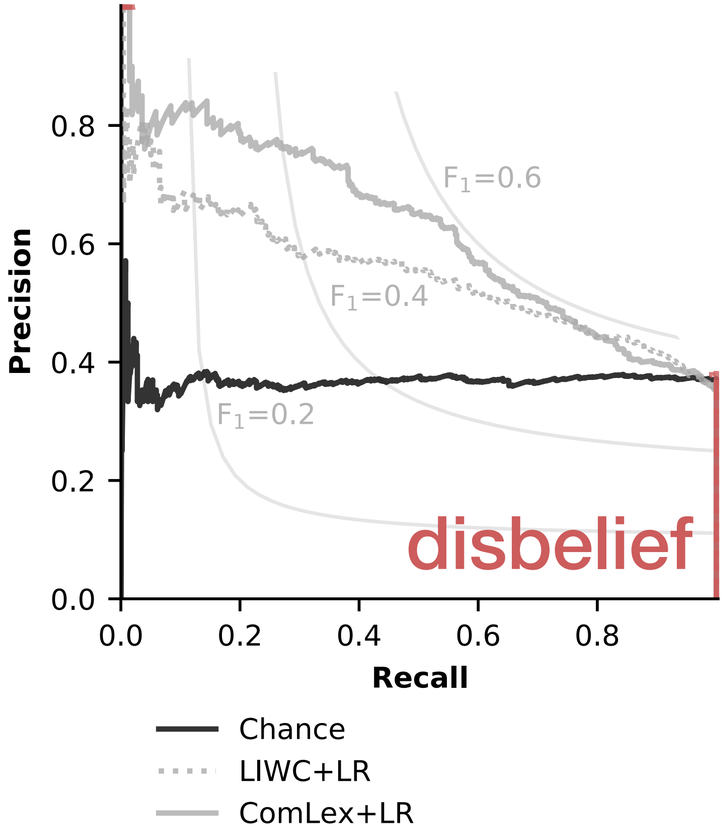


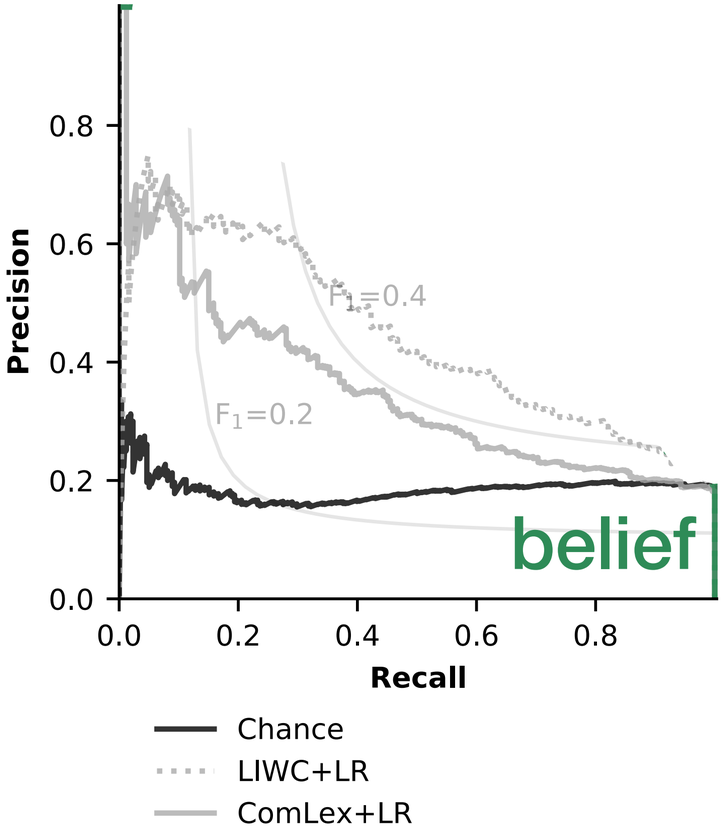
Model: lexicon feature + logistic regression

- LIWC / ComLex + logistic regression

Model: lexicon feature + logistic regression

- LIWC / ComLex + logistic regression



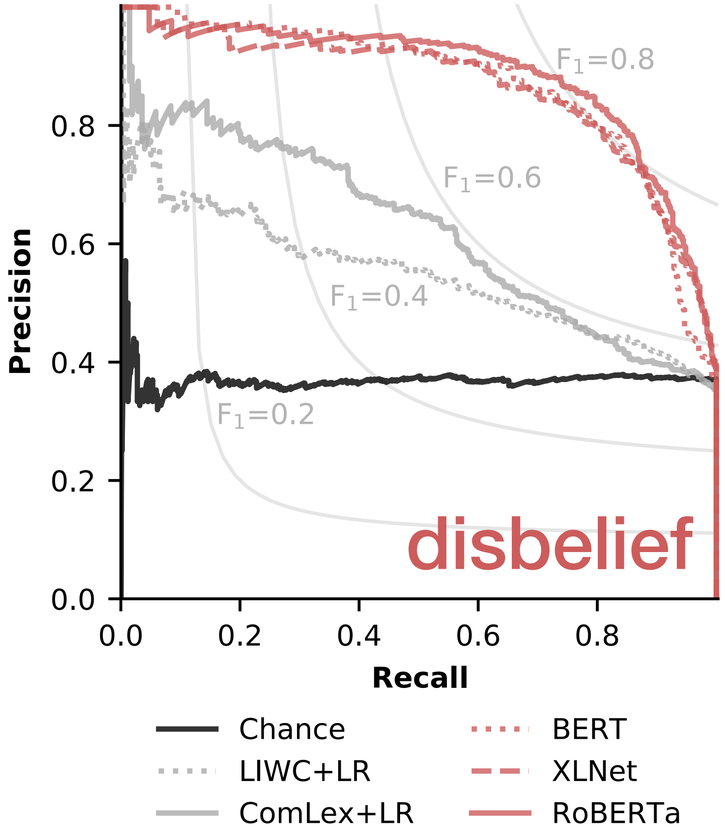


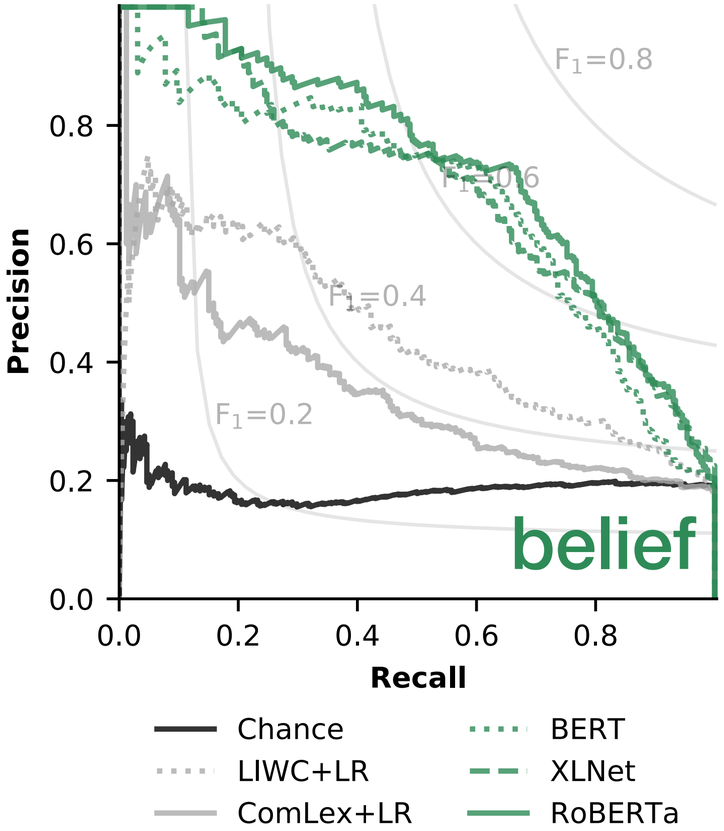
Model: sequence feature + transfer learning

- BERT / XLNet / RoBERTa

Model: sequence feature + transfer learning

- BERT / XLNet / RoBERTa





Model: tuning thresholds for unbiasedness

- Threshold τ : 1 if $\mathbb{P} > \tau$, 0 otherwise.

Model: tuning thresholds for unbiasedness

- Threshold τ : 1 if $\mathbb{P} > \tau$, 0 otherwise.
- Prevalence of (dis)belief b .

underlying prevalence

estimated prevalence

$$\mathbb{E}(b) = \mathbb{E}(\hat{b})$$

Model: tuning thresholds for unbiasedness

- Threshold τ : 1 if $\mathbb{P} > \tau$, 0 otherwise.
- Prevalence of (dis)belief b .

total *labeled* positive

$$\mathbb{E}(b) = \frac{TP(\tau) + FN(\tau)}{N}$$

sample size

Model: tuning thresholds for unbiasedness

- Threshold τ : 1 if $\mathbb{P} > \tau$, 0 otherwise.
- Prevalence of (dis)belief b .

total *predicted* positive

$$\frac{TP(\tau) + FP(\tau)}{N} = \mathbb{E}(\hat{b})$$

sample size

Model: tuning thresholds for unbiasedness

- Threshold τ : 1 if $\mathbb{P} > \tau$, 0 otherwise.
- Prevalence of (dis)belief b .

$$\mathbb{E}(b) = \frac{TP(\tau) + \boxed{FN(\tau)}}{N} = \frac{TP(\tau) + \boxed{FP(\tau)}}{N} = \mathbb{E}(\hat{b})$$

Model: tuning thresholds for unbiasedness

- Threshold τ : 1 if $\mathbb{P} > \tau$, 0 otherwise.
- Prevalence of (dis)belief b .

- Choose τ in the training set for $FN(\tau) = FP(\tau)$.
- Apply τ on the testing set.
- Test $FN(\tau) = FP(\tau)$ as null hypothesis.

Model: evaluation

Classifier	Disbelief					Belief				
	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1
Chance	0.654	✓	0.354	0.494	0.533	0.814	✓	0.170	0.490	0.691
LIWC+LR	0.415	✓	0.548	0.647	0.675	0.306	✓	0.450	0.666	0.806
ComLex+LR	0.364	✓	0.586	0.683	0.712	0.279	✓	0.371	0.612	0.761
BERT	0.374	✓	0.801	0.840	0.850	0.646	✗	0.620	0.773	0.877
XLNet	0.514	✓	0.798	0.839	0.850	0.593	✗	0.646	0.785	0.877
RoBERTa	0.436	✓	**0.817**	**0.855**	**0.864**	0.451	✓	**0.671**	**0.800**	**0.884**

Model: evaluation

Classifier	Disbelief					Belief				
	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1
Chance	0.654	✓	0.354	0.494	0.533	0.814	✓	0.170	0.490	0.691
LIWC+LR	0.415	✓	0.548	0.647	0.675	0.306	✓	0.450	0.666	0.806
ComLex+LR	0.364	✓	0.586	0.683	0.712	0.279	✓	0.371	0.612	0.761
BERT	0.374	✓	0.801	0.840	0.850	0.646	✗	0.620	0.773	0.877
XLNet	0.514	✓	0.798	0.839	0.850	0.593	✗	0.646	0.785	0.877
RoBERTa	0.436	✓	**0.817**	**0.855**	**0.864**	0.451	✓	**0.671**	**0.800**	**0.884**

- Highest F1 scores.

Model: evaluation

Classifier	Disbelief					Belief				
	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1
Chance	0.654	✓	0.354	0.494	0.533	0.814	✓	0.170	0.490	0.691
LIWC+LR	0.415	✓	0.548	0.647	0.675	0.306	✓	0.450	0.666	0.806
ComLex+LR	0.364	✓	0.586	0.683	0.712	0.279	✓	0.371	0.612	0.761
BERT	0.374	✓	0.801	0.840	0.850	0.646	✗	0.620	0.773	0.877
XLNet	0.514	✓	0.798	0.839	0.850	0.593	✗	0.646	0.785	0.877
RoBERTa	0.436	✓	**0.817**	**0.855**	**0.864**	0.451	✓	**0.671**	**0.800**	**0.884**

- Highest F1 scores.
- Unbiased (p<0.01 w/ Bonferroni correction).

Measurement: dataset

the entire archive of PolitiFact & Snopes

- Fact-checks from ~~PolitiFact between Jan to Jun 2019~~, whose claims were originated from ~~Twitter~~.

Twitter, Facebook & YouTube

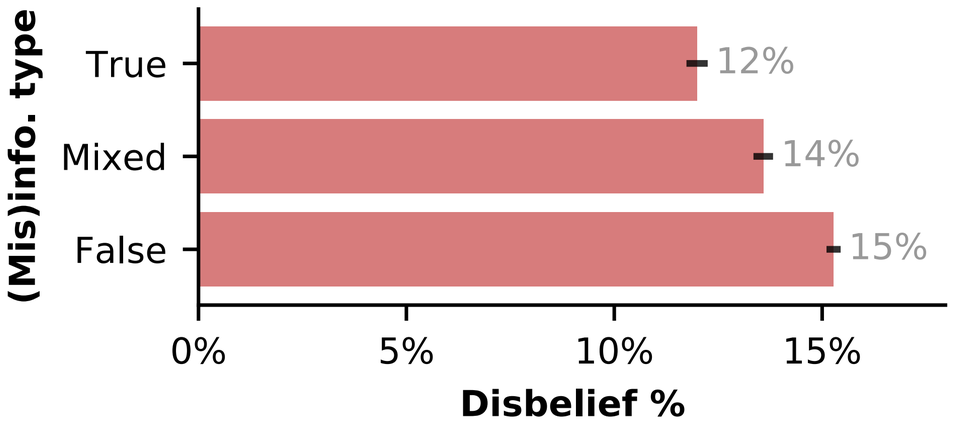
[3] Jiang, S., and Wilson, C. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *PACMHCI (CSCW)*.

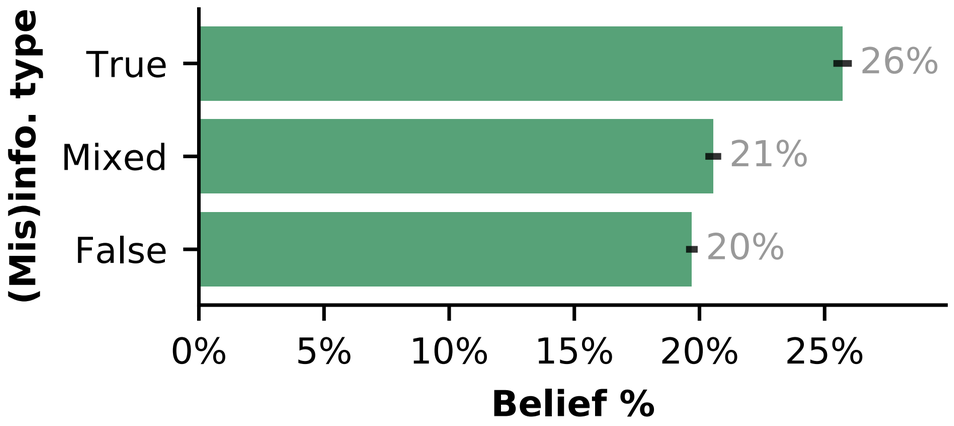
Measurement: dataset

- Fact-checks from ~~PolitiFact between Jan to Jun 2019~~ the entire archive of PolitiFact & Snopes, whose claims were originated from ~~Twitter.~~ Twitter, Facebook & YouTube
- Query APIs and find all comments to the claim.
- ~~18~~ 5,303 claims, ~~6,809~~ 2,614,374 comments.

[3] Jiang, S., and Wilson, C. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *PACMHCI (CSCW)*.

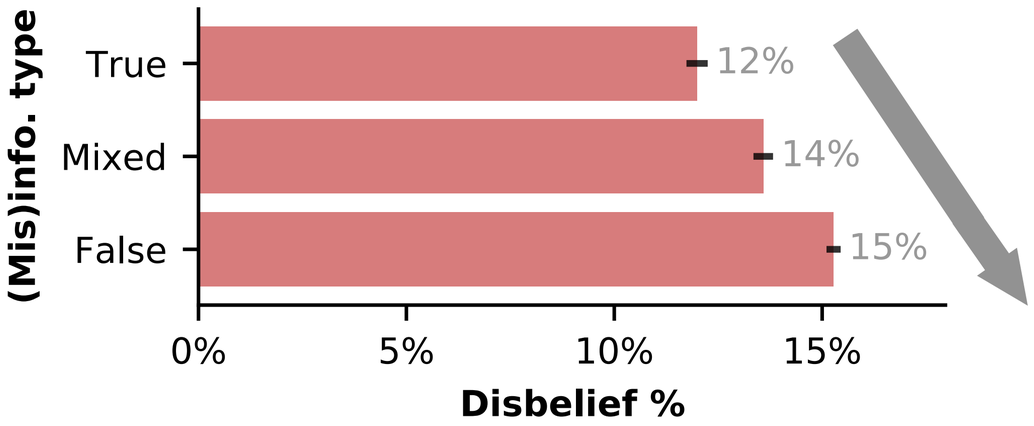
Measurement: overall prevalence

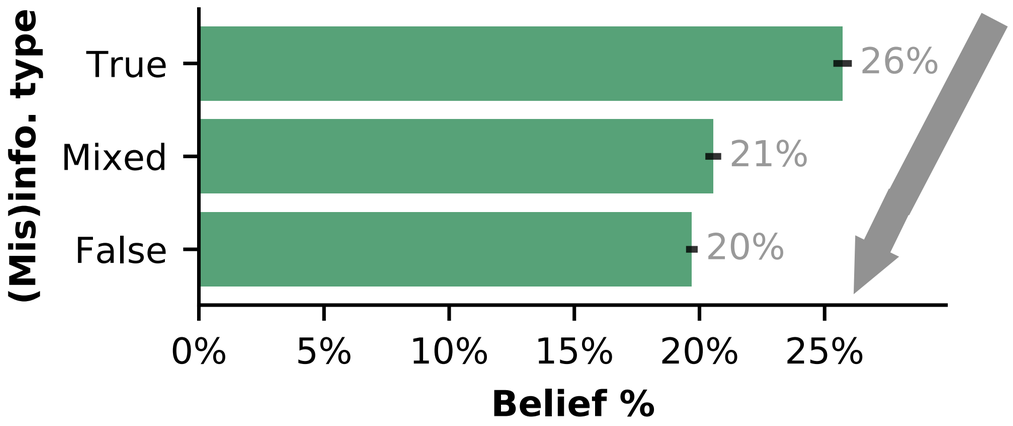




- 12% - 15% of disbelief, 20% - 26% of belief.

Measurement: overall prevalence





- 12% - 15% of disbelief, 20% - 26% of belief.
- With veracity decreasing, disbelief increases and belief decreases.

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

prevalence of (dis)belief (%)

$$\boxed{\hat{b}} = \beta_0 + \beta_1 \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

prevalence of (dis)belief (%)

$$\hat{b} = \beta_0 + \beta_1 \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

time difference between the comment and its claim (day)

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

prevalence of (dis)belief (%)

time difference between the comment and its fact-check (day)

$$\hat{b} = \beta_0 + \beta_1 \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

time difference between the comment and its claim (day)

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

prevalence of (dis)belief (%)

if a comment was posted before/after fact-check (1)

$$\hat{b} = \beta_0 + \beta_1 \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

time difference between the comment and its claim (day)

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

initial prevalence (%)

$$\hat{b} = \boxed{\beta_0} + \beta_1 \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

initial prevalence (%)

$$\hat{b} = \boxed{\beta_0} + \boxed{\beta_1} \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

the effect of time (%/day)

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.

initial prevalence (%) effect of fact-checking (%)

$$\hat{b} = \boxed{\beta_0} + \boxed{\beta_1} \cdot \Delta_{C_m,m} + \boxed{\beta_2} \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

the effect of time (%/day)

Measurement: effects of time and fact-checking

- Effects of time and fact-checking for false claims.
- Prevalence of (dis)belief b.
- OLS for model estimation.

$$\hat{b} = \beta_0 + \beta_1 \cdot \Delta_{C_m,m} + \beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m}) + \epsilon$$

Measurement: effects of time and fact-checking

Parameters	Disbelief		Belief	
	Estimation	p-value	Estimation	p-value
$\hat{\beta}_0$	$+1.52 \times 10^{-1}$	***	$+1.98 \times 10^{-1}$	***
$\hat{\beta}_1$	$+9.96 \times 10^{-6}$	***	-2.19×10^{-5}	***
$\hat{\beta}_2$	$+5.00 \times 10^{-2}$	***	-3.41×10^{-2}	***
# of samples	$1,395,293$		$1,395,293$	

Measurement: effects of time and fact-checking

Parameters	Disbelief		Belief	
	Estimation	p-value	Estimation	p-value
$\hat{\beta}_0$	$+1.52 \times 10^{-1}$	***	$+1.98 \times 10^{-1}$	***
$\hat{\beta}_1$	$+9.96 \times 10^{-6}$	***	-2.19×10^{-5}	***
$\hat{\beta}_2$	$+5.00 \times 10^{-2}$	***	-3.41×10^{-2}	***
# of samples	$1,395,293$		$1,395,293$	

For false claims:

- disbelief increases 0.001%/day.
- belief decreases 0.002%/day.

Measurement: effects of time and fact-checking

Parameters	Disbelief		Belief	
	Estimation	p-value	Estimation	p-value
$\hat{\beta}_0$	$+1.52 \times 10^{-1}$	***	$+1.98 \times 10^{-1}$	***
$\hat{\beta}_1$	$+9.96 \times 10^{-6}$	***	-2.19×10^{-5}	***
$\hat{\beta}_2$	$+5.00 \times 10^{-2}$	***	-3.41×10^{-2}	***
# of samples	$1,395,293$		$1,395,293$	

- disbelief increases 0.001%/day.
- belief decreases 0.002%/day.

For false claims:

- disbelief increases 5% after fact-check.
- belief decreases 3.4% after fact-check.

Discussion: limitations

- Topical bias: mostly political issues.

Discussion: limitations

- Topical bias: mostly political issues.
- Proxy validity: modeling only *expressed* (dis)belief.

Thank you!

Please send questions to: sjiang@ccs.neu.edu

Data and code available at: misinfo.shanjiang.me

Northeastern University

UC SANTA BARBARA