Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media

Shan Jiang and Christo Wilson

Northeastern University

Outline

- Background
- Data
- Methods
- Results
- Discussion

Background: Misinformation - What is it?

Mis · information

wrong, incorrect, inaccurate, etc.

Background: Misinformation - Example

An ‘extremely credible source’ has called my office and told me that @BarackObama’s birth certificate is a fraud.

Background: Information Veracity

Background: Fact-Checking - Example

Snopes

POLITIFACT

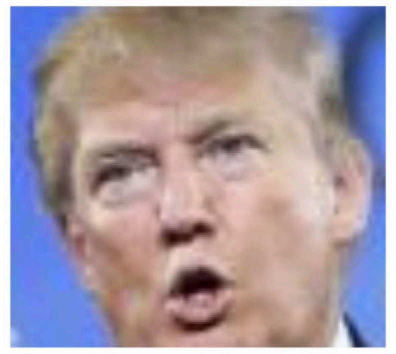
CLAIM

Barack Obama's birth certificate is a forgery. See Example(s)

RATING

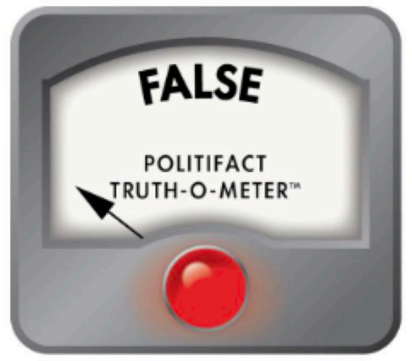
FALSE

Share The Facts



Donald Trump

Presidential candidate



POLITIFACT

Says President Obama's "grandmother in Kenya said he was born in Kenya and she was there and witnessed the birth."

In an interview. – Thursday, April 7, 2011

SHARE READ MORE

Background: User Comments - Example

An 'extremely credible source' has called my office and told me that @BarackObama's birth certificate is a fraud.

An 'extremely credible source' has told me that Donald Trump's presidency is a fraud.

Lol. Sure. Was it someone from Fox News?

https://www.snopes.com/fact-check/birth-certificate

Background: User Comments - Linguistic Signals

An ‘extremely credible source’ has called my office and told me that @BarackObama’s birth certificate is a fraud.

An ‘extremely credible source’ has told me that Donald Trump’s presidency is a fraud.

Fake

Lol. Sure. Was it someone from Fox News?

Laughter

https://www.snopes.com/fact-check/birth-certificate

Fact

Background: Research Goal

Systematically analyze such linguistic signals.

Background: Why Do We Care? - Misinformation

Macro (previous work):

Changed outcome of the 2016 US presidential election?

— Probably NO (Allcott et al. 2017; Guess et al 2018).

Micro (our focus):

Twisted facts → Reduce trust?

Inflammatory language → Discourage reasoned conservation?

Signals in user comments indicating these effects?

Background: Why Do We Care? - Fact-Checking

Corrective effect:
Changing people's beliefs (Fridkin et al. 2015; Porter et al. 2018).

"Backfire" effect:
Leaning stronger to false beliefs (Wood et al. 2016; Haglin et al. 2018).

Signals in user comments indicating these effects?

Background: Research Questions

Do users exhibit different linguistic signals —

RQ1). — under posts with different veracity (from true to false)?
RQ1a). Misinformation-awareness signals?
RQ1b). Emotional and topical signals?

RQ2). — before and after a post is fact-checked?
RQ2a). Signals indicating corrective effect?
RQ2b). Signals indicating “backfire" effect?

Outline

- ~~Background~~
- Data
- Methods
- Results
- Discussion

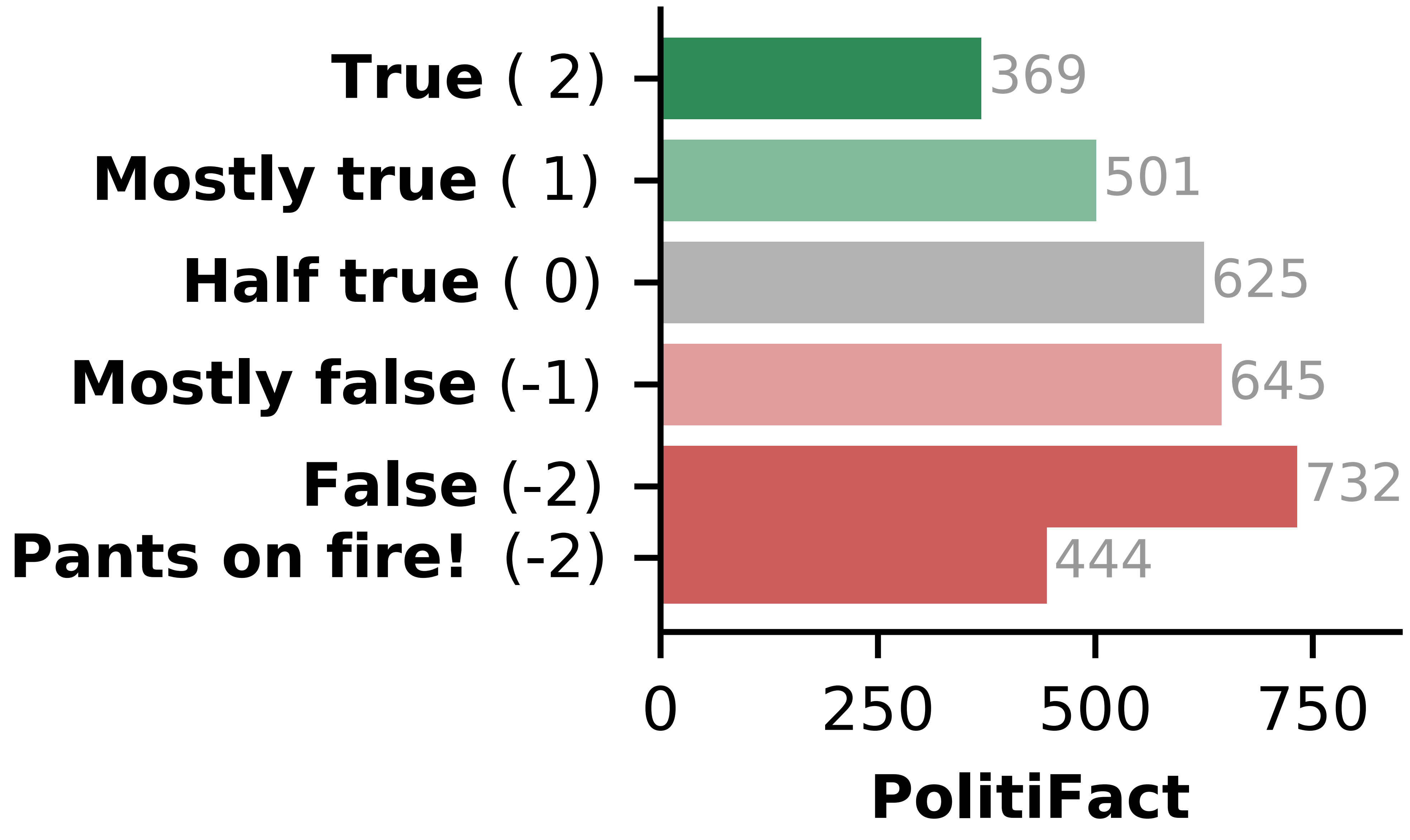
Data: Framework

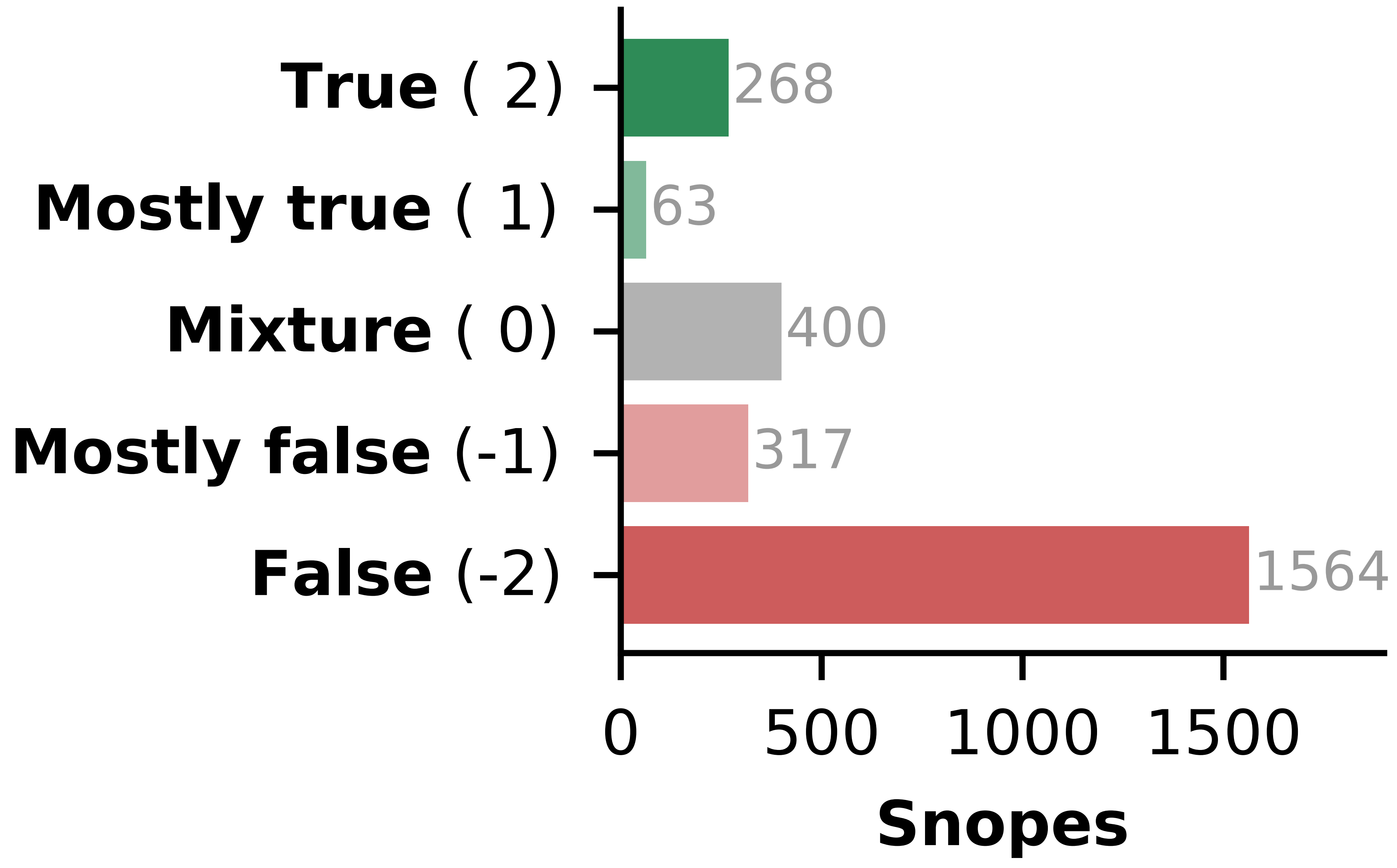
Fact-check articles → Social media posts → User comments

Data: Fact-Check Articles

Fact-check articles → Social media posts → User comments

14,184 articles from POLITIFACT, 11,345 from Snopes;
5,303 were sourced from YouTube, Facebook or Twitter.

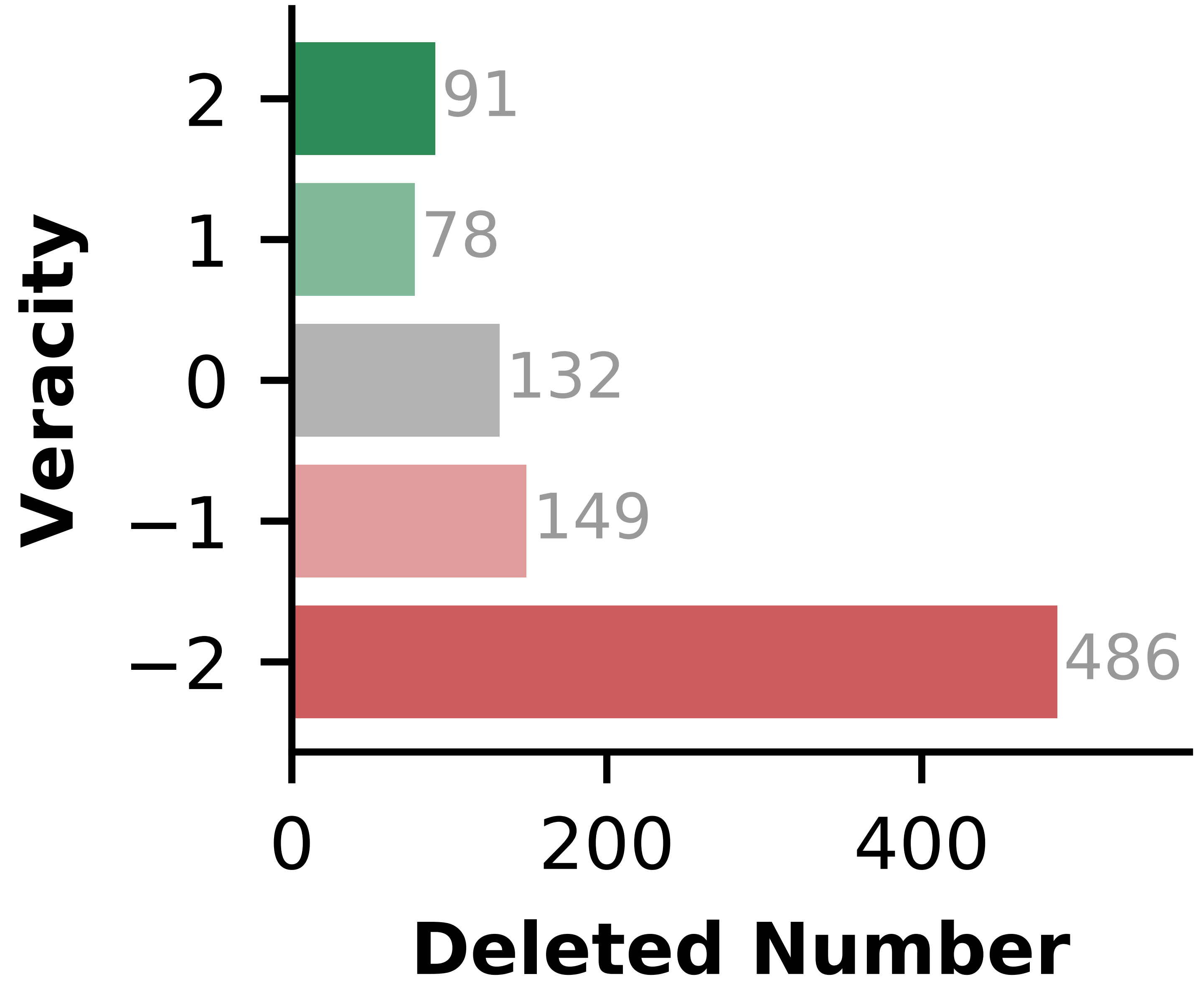




Data: Social Media Posts

Fact-check articles → **Social media posts** → User comments

31% posts were deleted;
82% had veracity <= 0 (mostly false or false).



Data: User Comments

Fact-check articles → Social media posts → User comments

1,672,687 from Facebook ;

113,687 from Twitter ;

828,000 from YouTube .

Outline

- ~~Background~~
- ~~Data~~
- Methods
- Results
- Discussion

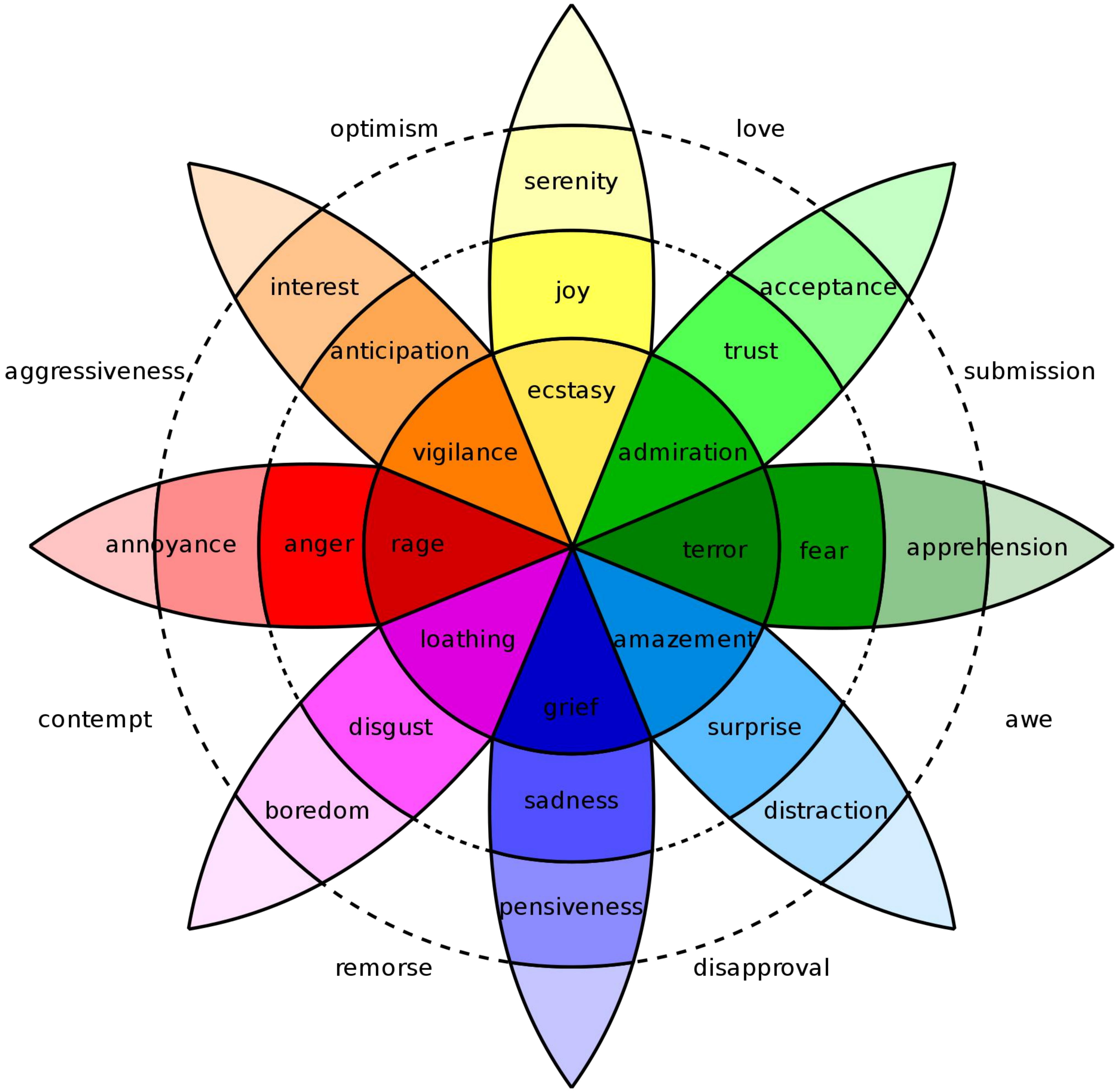
Methods: How?

Existing lexicons:

EmoLex
(Plutchik's wheel of emotions)

LIWC
(most extensively used)

……



Methods: Problem

Existing lexicons:

EmoLex
(Plutchik's wheel of emotions)

LIWC
(most extensively used)

……

Not context-specific:

No emojis;
No fake / fact clusters;
Limited set of swear words;
……

Methods: **ComLex**

A new context-specific lexicon: **ComLex**

Methods: **Words**

pew

snopes

politifact

fact

hoax

scam

conspiracy

💩

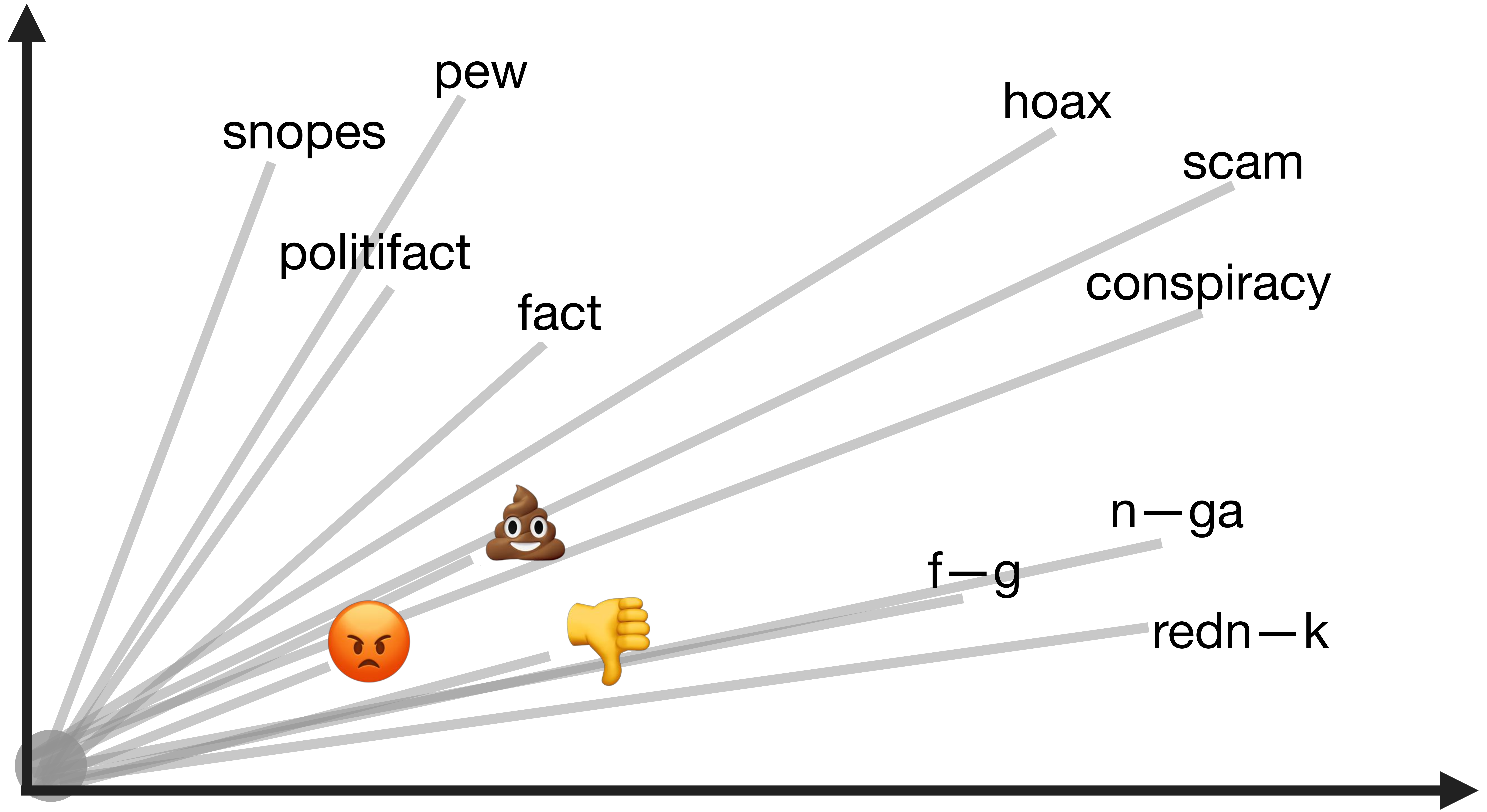
😡 👎

n—ga

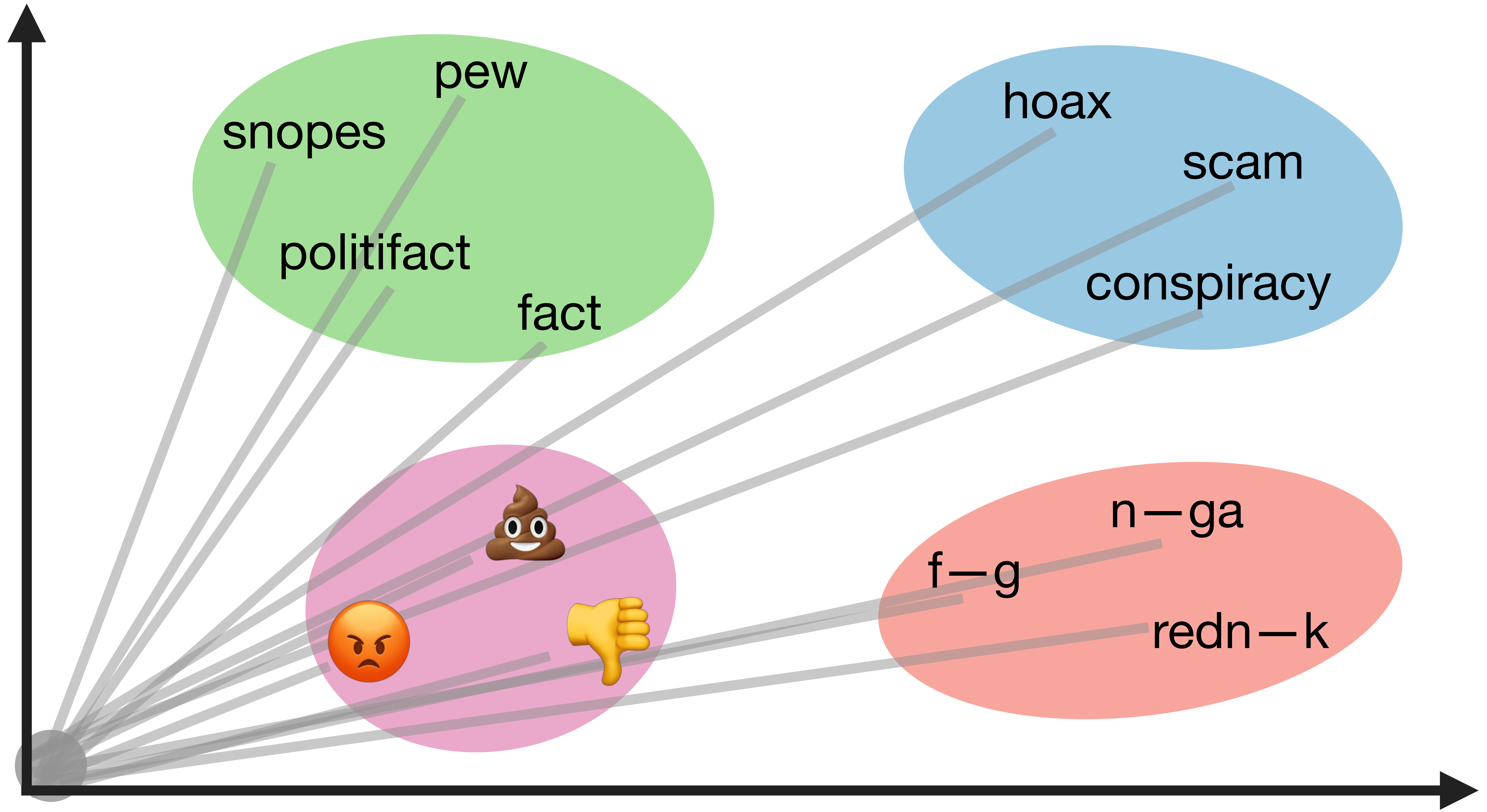
f—g

redn—k

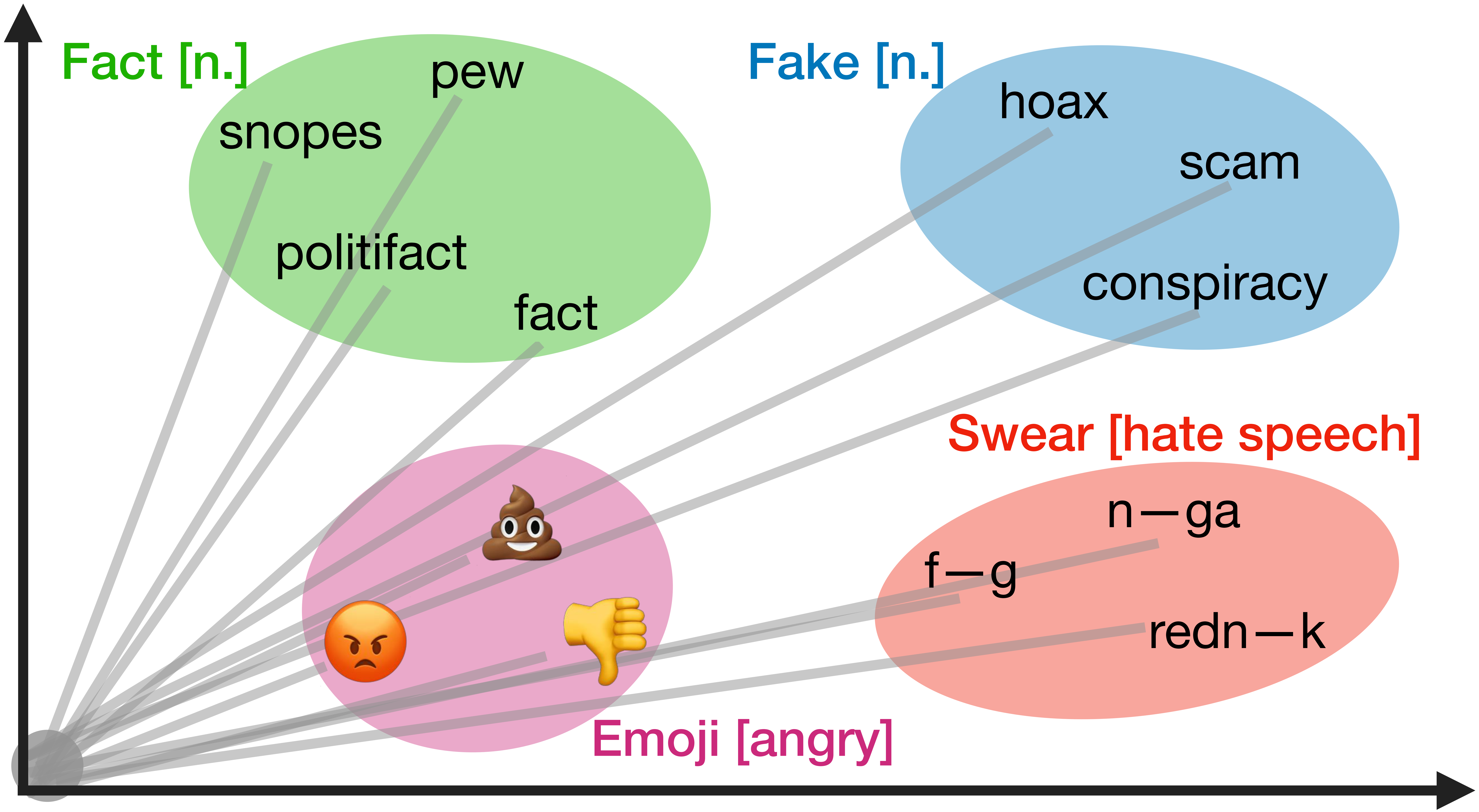
Methods: Word2Vec



Methods: Clustering



Methods: Manual Labeling



Methods: Validation

Manually selected **56** clusters → **ComLex.**

1). Human evaluation;
2). ComLex with LIWC (Pennebaker et al. 2015) and Empath (Fast et al. 2016);
3). Application and generalization (Pang et al. 2002; Ott et al. 2011).

(More details in our paper.)

Outline

- ~~Background~~
- ~~Data~~
- ~~Methods~~
- Results
- Discussion

Results: RQ1) Misinformation

Do users exhibit different linguistic signals
under posts with different veracity (from true to false)?

Results: Misinformation - Awareness

From true to fake, more likely to be aware of misinformation:

Fake [v. & adj.] (*fake, mislead, fabricate, …*)
Fake [n., bias] (*propaganda, rumor, distortion, …*)
Fake [n., false] (*hoax, scam, conspiracy, …*)
e.g., “this is fake news”, “this is brainwash propaganda”

From true to fake, trust decreases:

Trust [EmoLex] (*accountable, lawful, scientific, …*)

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: Misinformation - Emojis

From true to fake, emoji usage increases:

- Emoji [gesture] (🙏, 👋, 👊, …)
- Emoji [laughter] (😂, 🤣, 🙄, …)
- Emoji [happiness] (😃, 😊, 😍, …)
- Emoji [doubt] (❓, 🤷‍♀️, 🤷‍♂️, …)
- Emoji [sadness] (💔, 😢, 😭, …)
- Emoji [surprise] (😱, 😳, 😲, …)
- Emoji [anger] (👎, 😡, 💩, …)

e.g., “so ridiculous 😂😂”, “really? 😭”, “i smell bull 💩”.

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: Misinformation - Swear

From true to fake, swear increases:

Swear	[informal, common]	(*fuck, fuckin, damn, …*)
Swear	[informal, moderated]	(**, sh, fu, …*)
Swear	[hate speech]	(*n—ga, f—g, redn—k, …*)
Swear	[informal, other]	(*bastard, fucktards, …*)
Swear	[LIWC]	(*fuck, fu, dumbfuck, …*)
Swear	[informal, belittling]	(*moron, fool, loser, …*)

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: Misinformation - Objectivity

From true to fake, subjectivity increase:

Superlative [compare] (*dumbest, smartest, craziest, …*)

e.g., "dumbest thing i've seen today".

From true to fake, objectivity decreases:

Causal [LIWC] (*why, because, therefore, …*)

Comparative [compare] (*better, bigger, harder, …*)

e.g., "she would do better".

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: Misinformation - Topics

From true to fake, less likely to discuss concrete topics:

Work	[LIWC]	(*work, earn, payroll, …*)
Financial	[economy]	(*bill, budget, policy, …*)
Money	[LIWC]	(*financially, worth, income, …*)
Power	[LIWC]	(*worship, command, mighty, …*)
Financial	[monetary]	(*money, tax, dollar, …*)
Reward	[LIWC]	(*promotion, award, success, …*)
Society	[civilian]	(*people, public, worker, …*)
Achieve	[LIWC]	(*award, honor, prize, …*)
Health	[insurance]	(*health, healthcare, …*)
Admin	[n.]	(*attorney, secretary, …*)

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: RQ2) Fact-Checking

Do users exhibit different linguistic signals **before** and **after** a post is fact-checked?

Results: Fact-Checking - Corrective

After fact-checking, more likely to be aware of misinformation:

Fact	[n.]	(*fact, evidence, data, …*)
Fake	[v. & adj.]	(*fake, mislead, fabricate, …*)

After fact-checking, less doubtful emojis:

Emoji	[doubt]	(❓, 🤷‍♀️, 🤷‍♂️, …)

Note: Less significant - limited corrective effect?

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: Fact-Checking - “Backfire”

After fact-checking, more likely to use swear words:

Swear [informal, common] (*fuck, fuckin, damn, …*)

Note: Less significant, only 1 swear cluster - limited “backfire” effect?

*** $p < 0.001$
** $p < 0.01$
* $p < 0.05$

Results: "Backfire" - Example

"Obamacare" mandates that no one over 75 will be given major medical procedures unless approved by an ethics panel.

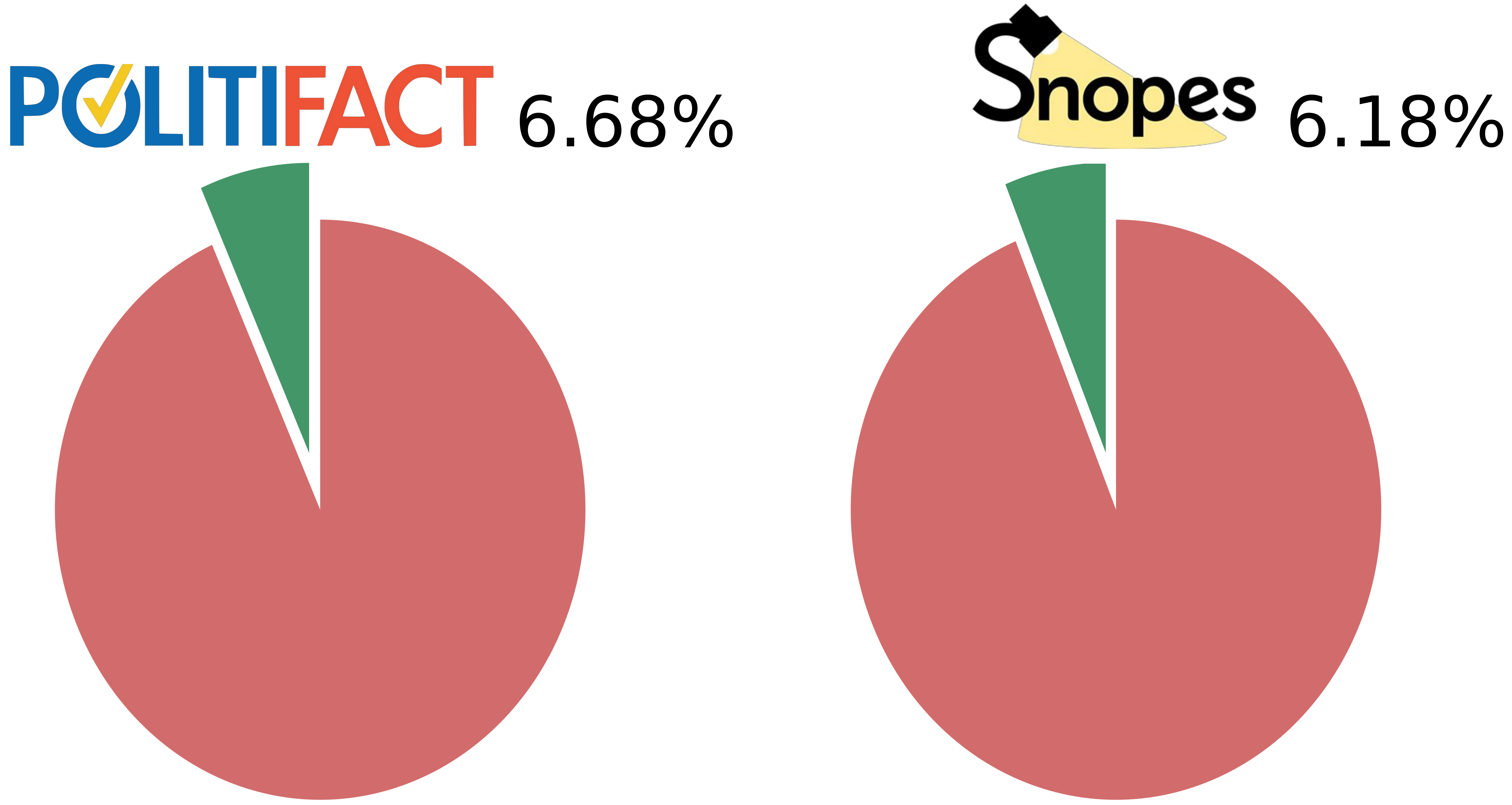
FALSE

What it has to do with moron is you are a fucking sheep that believes everything he's told… and everything you see that doesn't fit what you want to hear is "propaganda" then you mention snopes and politifact like its the official source for truth… Youtube videos are a pretty good source for truth…

View fact-checking itself as biased in general, rather than "backfire" because of individual fact-check articles.

Results: Fact-Checking - Reference

Only ~**6%** fact-checked posts have comments referring to the fact-check article.



Outline

- ~~Background~~
- ~~Data~~
- ~~Methods~~
- ~~Results~~
- Discussion

Discussion: Application

Misinformation detection:
(Spearman correlation)

Random guess:	0.000
Neural Nets using **EmoLex**:	0.081
Neural Nets using **LIWC**:	0.116
Neural Nets using **ComLex**:	0.214
Maximum:	1.000

→ **Not good.**

(More details in our paper.)

Discussion: Application

Misinformation detection:
(Spearman correlation)

Random guess:	0.000
Neural Nets using **EmoLex**:	0.081
Neural Nets using **LIWC**:	0.116
Neural Nets using **ComLex**:	0.214
Maximum:	1.000

Text + Meta: **0.20-0.27**
(Yang 2017)

Not good.

(More details in our paper.)

Discussion: Limitations

Only focus on social media. → Other news sources?

Only focus on veracity. → Intentionality?

Discussion: Takeaways

Users do exhibit different linguistic signals —

RQ1). — under posts with different veracity (from true to false).
RQ1a). More misinformation-awareness signals.
RQ1b). More emojis, more swear words, less concrete topics, etc.

RQ2). — after a post is fact-checked.
RQ2a). Some signals indicating corrective effect.
RQ2b). Some signals indicating “backfire" effect.

Data & lexicon & code available at: misinfo.shanjiang.me
A Blog highlighting findings available on **CSCW Medium**

Thanks!

Shan Jiang
Email: sjiang@ccs.neu.edu

Northeastern University

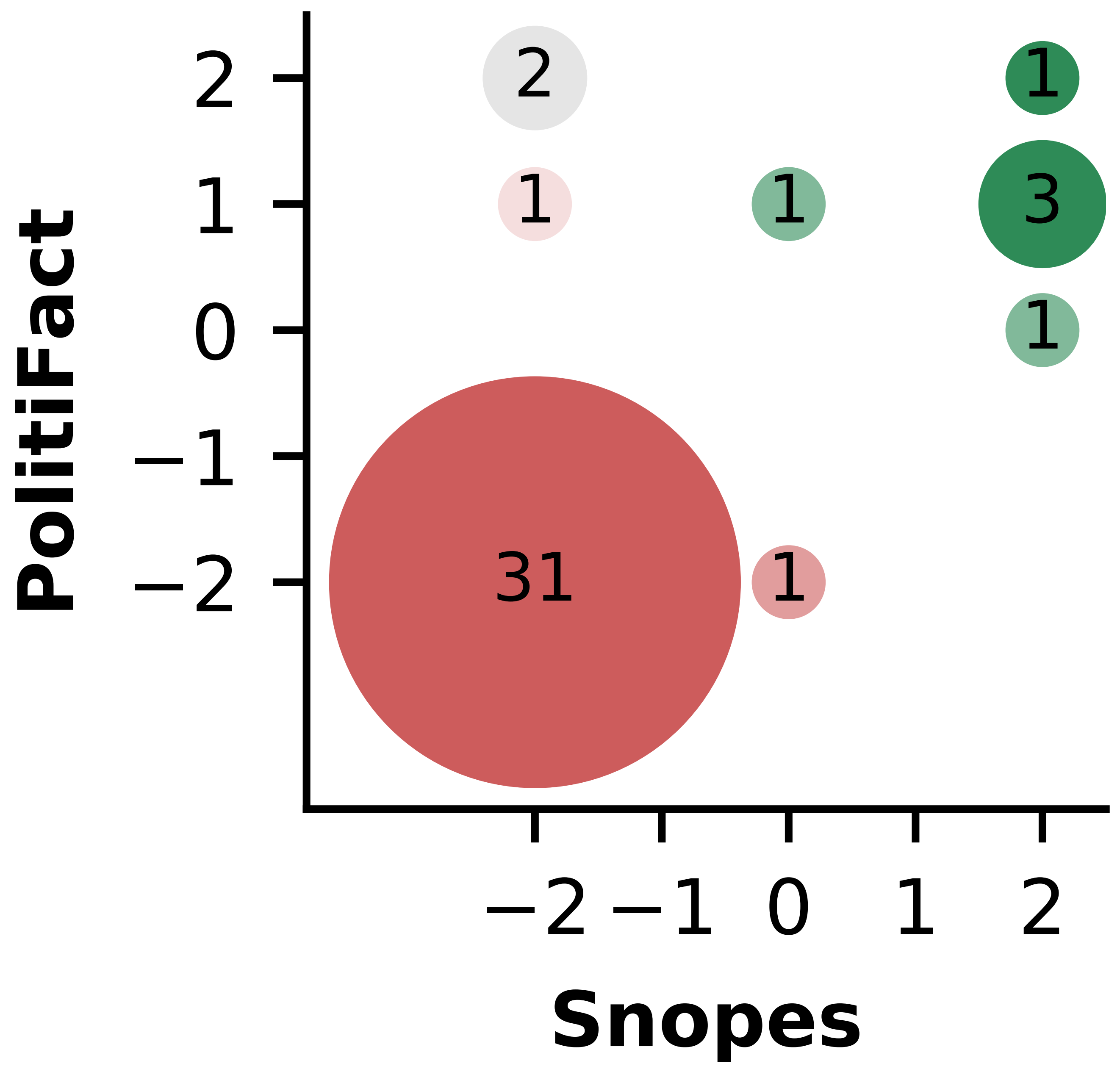
References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-36.
Guess, A., Nyhan, B., & Reifler, J. (2018). Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. European Research Council.
Fridkin, K., Kenney, P. J., & Wintersieck, A. (2015). Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising. Political Communication, 32(1), 127-151.
Porter, E., Wood, T. J., & Kirby, D. (2018). Sex trafficking, Russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news. Journal of Experimental Political Science, 5(2), 159-164.
Wood, T., & Porter, E. (2016). The elusive backfire effect: Mass attitudes' steadfast factual adherence. Political Behavior, 1-29.
Haglin, K. (2017). The limitations of the backfire effect. Research & Politics, 4(3), 2053168017716547.
Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015.
Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4647-4657). ACM.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL.
Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of ACL-HLT.
Wang, W. Y. (2017). " Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proceedings of ACL.

Data: Fact-Check Agreement

Fact-check articles → Social media posts → User comments

41 posts are fact-checked by both POLITIFACT and Snopes;
They highly agree with each other (Spearman $r = 0.671^{***}$).

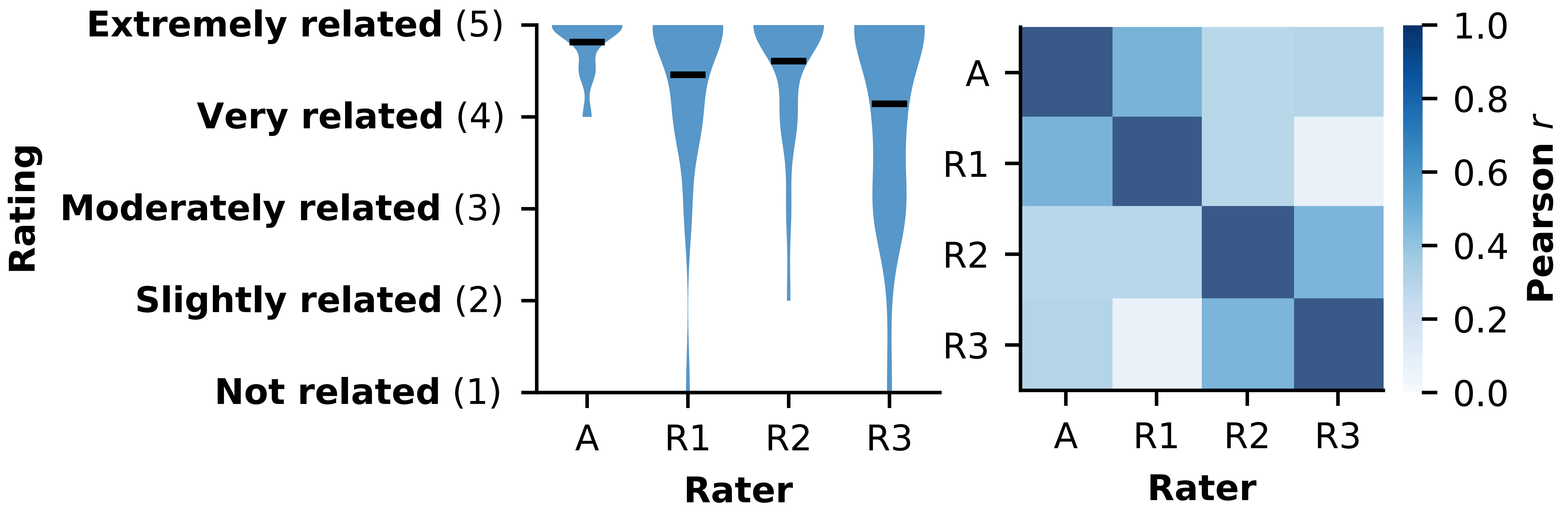


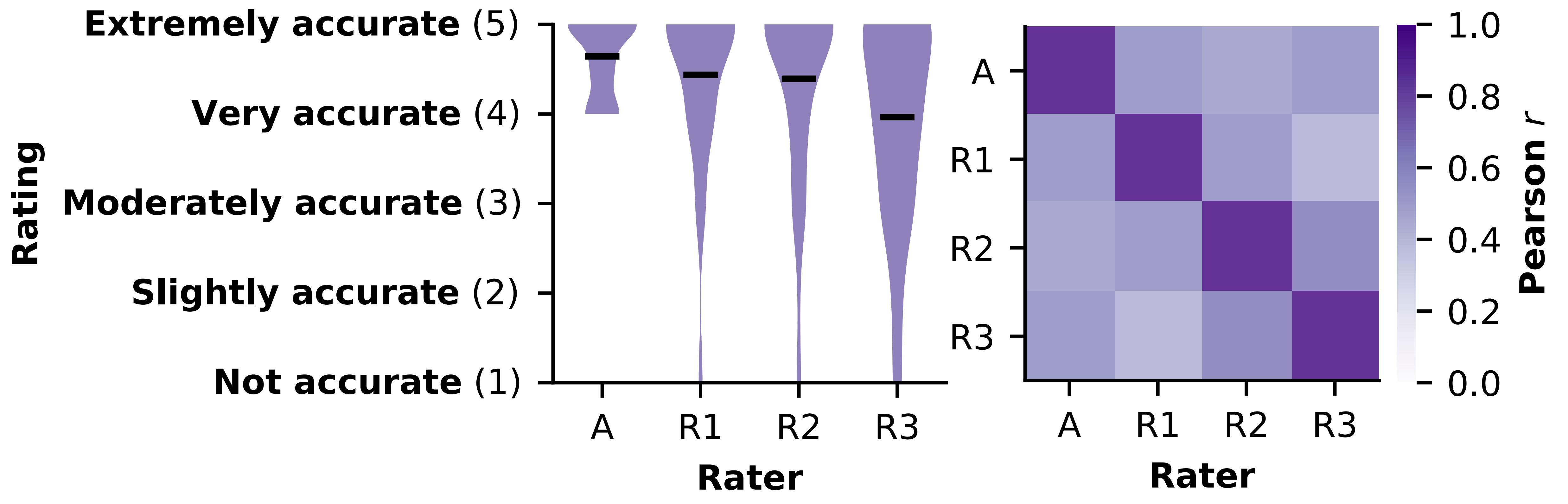
Evaluation: Human Evaluation

Human Evaluation → Comparing with Others → Application

Rating 1: Semantic Closeness (1-5), mean: **4.506**

Rating 2: Labeling Accuracy (1-5), mean: **4.359**





Evaluation: Comparing with Others

Human Evaluation → Comparing with Others → Application

Similar Cluster with LIWC:
Family (Pearson r = 0.883***).
Pronoun (Pearson r = 0.877***).
Preposition (Pearson r = 0.833***).

Similar Cluster with Empath:
Monster (Pearson r = 0.949***).
Timidity (Pearson r = 0.904***).
Ugliness (Pearson r = 0.908***).

Evaluation: Application

Human Evaluation → Comparing with Others → Application

Dataset	Lexicon	Model	Accuracy*
Hotel reviews (Ott et al. 2011)	Human judges		56.9% - 61.9%
	GI	SVM	73.0%
	LIWC	SVM	76.8%
	ComLex	SVM	**81.4%**
	Learned unigrams	SVM	88.4%
Movie reviews (Pang et al. 2002)	Human judges		58.0% - 69.0%
	ComLex	SVM	**72.3%**
	Learned unigrams	SVM	72.8%