Modeling Intergroup Bias in Online Conversation

PH.D DEFENSE

Venkat

April 12, 2024

The University of Texas at Austin

How is **in-group** speech different from **out-group** speech?

Most work in NLP approaches bias as negative or perjorative language use towards an individual or group based on demographics.

Most work in NLP approaches bias as negative or perjorative language use towards an individual or group based on demographics.

However, research in psychology and social science suggests that bias is difference in behavior situated in relationships between people, and context. **All language use is biased**.

How do we bring this insight into our work?

The LIB hypothesis tries to explain the persistence of stereotypes through systematic language variation between **in-group** and **out-group** language.

The LIB hypothesis tries to explain the persistence of stereotypes through systematic language variation between **in-group** and **out-group** language.

LIB hypothesizes that abstract predicates are used when a description **conforms to stereotype**.

(1)
a. The man police want to talk to probably **hit** the victims.
b. The man police want to talk to probably **hurt** the victims.
c. The man police want to talk to probably **hated** the victims.
d. The man police want to talk to is probably **violent**.

We can study systematic differences in interpersonal language *inspired by the LIB*, and this can be an **effective framing** of social bias — intergroup bias.

① Intergroup bias can be analyzed through decomposition into **relationship** (in-group vs. out-group) and **emotion** in political tweets.

① Intergroup bias can be analyzed through decomposition into **relationship** (in-group vs. out-group) and **emotion** in political tweets.

② By grounding intergroup bias in a robust description of events preceding an utterance, we find that **form of referent varies linearly** with the grounded descriptions.

1. Intergroup bias in political tweets.

1. Intergroup bias in political tweets.
2. Counterfactual probing for intergroup bias.

① Intergroup bias in political tweets.

② Counterfactual probing for intergroup bias.

③ Grounding intergroup bias in football comments.

② a. **Admire** Chairman @reprichmond's moral voice on issues of racism and restorative justice. He is **a real leader** for our nation and Congress.

b. Parents and families live in constant fear for their children with food allergies. A worthy **bipartisan** cause - thank you @drphilroe for your **leadership** on this issue.

② a. **Admire** Chairman @reprichmond's moral voice on issues of racism and restorative justice. He is **a real leader** for our nation and Congress.

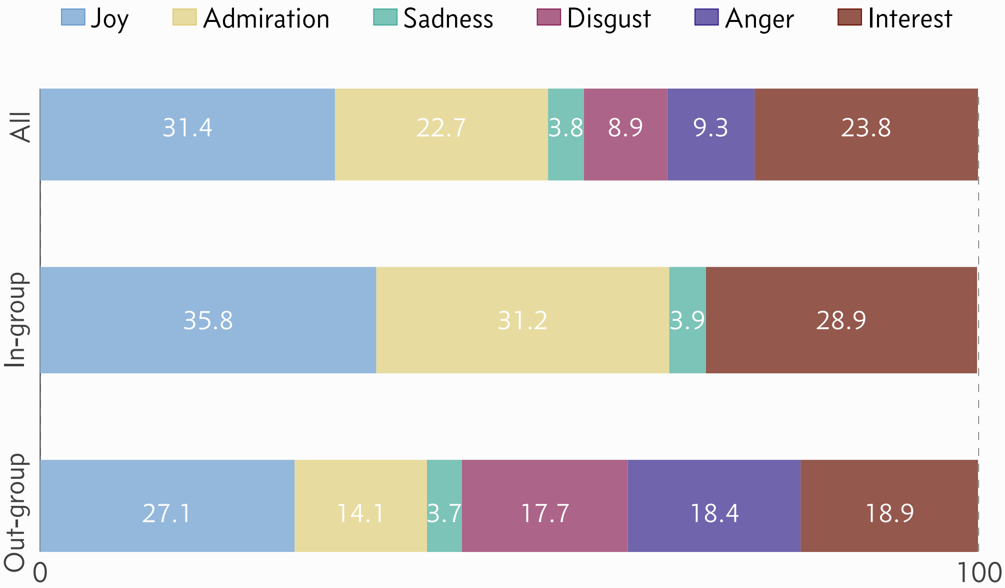
b. Parents and families live in constant fear for their children with food allergies. A worthy **bipartisan** cause - thank you @drphilroe for your **leadership** on this issue.

These utterances differ along two **interpersonal** dimensions:

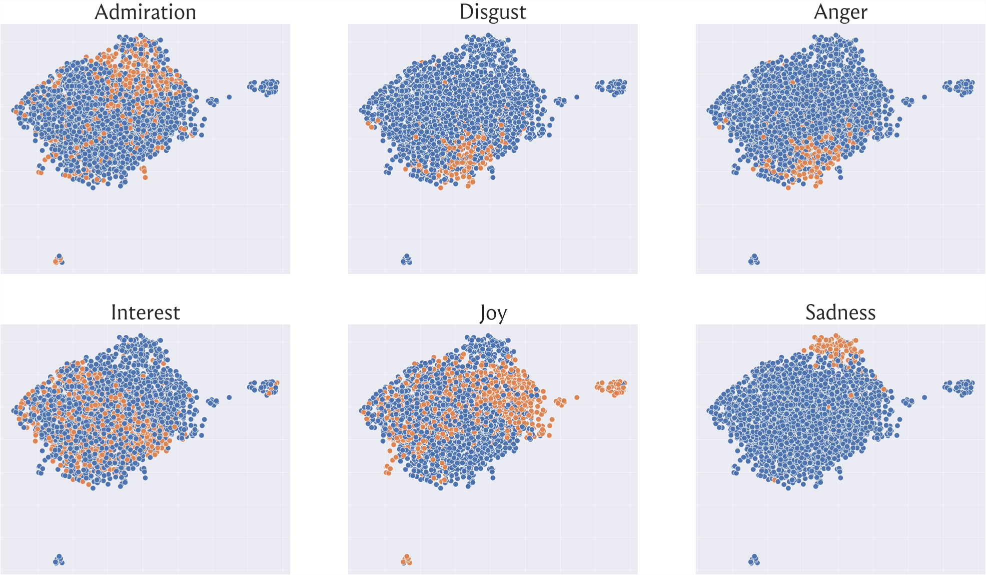
- the relationship between speaker and target — (a) is **in-group**, (b) is **out-group**.
- emotion expressed by speaker towards target.

EMOTION DISTRIBUTION

EMOTION DISTRIBUTION

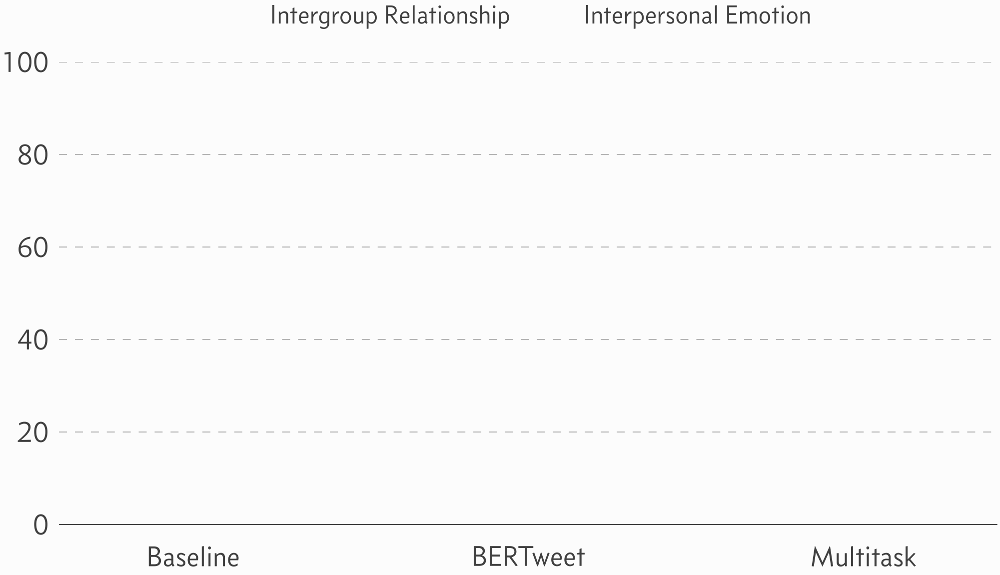


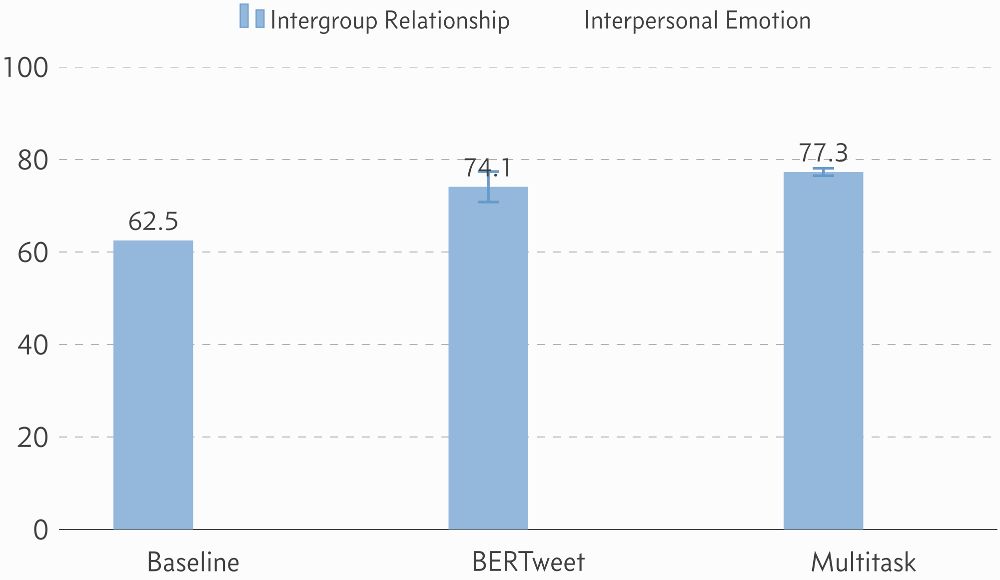
TWEET EMBEDDINGS & GOLD EMOTIONS

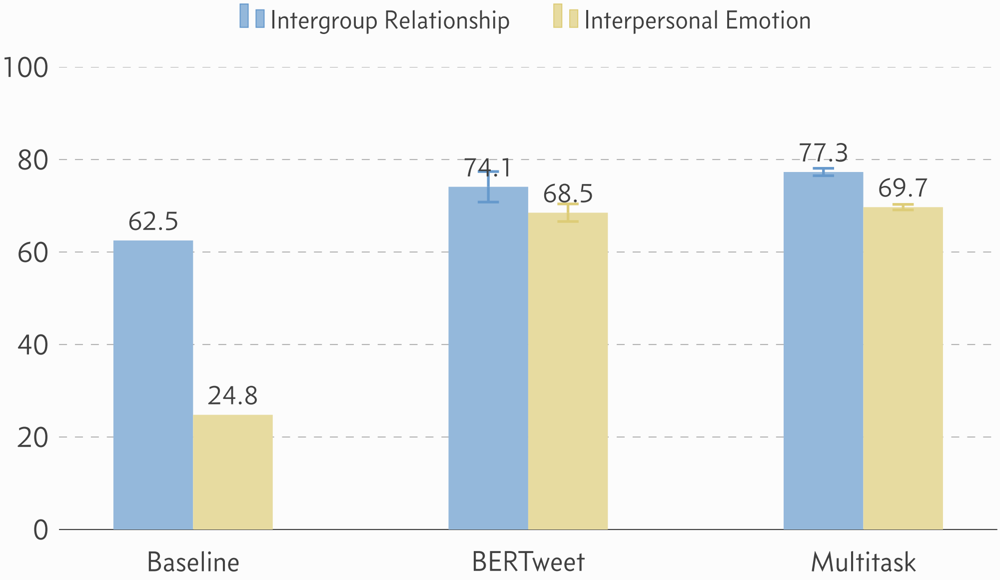


Tweet embeddings from a language model projected downward to 2 dimensions. Each point is a tweet and orange indicates the emotion is present. Observe the separability of clusters of emotions.

RESULTS


Intergroup Relationship
Interpersonal Emotion
100
80
60
40
20
0
Baseline
BERTweet
Multitask


Intergroup Relationship
Interpersonal Emotion
100
80
60
40
20
0
62.5
74.1
77.3
Baseline
BERTweet
Multitask



Recall score. Multitasking improves on vanilla model (slightly).

- Interpersonal emotion and intergroup relationship, two dimensions of intergroup bias, co-vary systematically.

- Interpersonal emotion and intergroup relationship, two dimensions of intergroup bias, co-vary systematically.
- Multitask modeling provides further evidence that the two are intertwined.

- Interpersonal emotion and intergroup relationship, two dimensions of intergroup bias, co-vary systematically.
- Multitask modeling provides further evidence that the two are intertwined.
- What is the actual **linguistic variation**? How does it interact with **situational context**?

	In-group	**Out-group**
socially desirable	abstract	concrete
socially undesirable	concrete	abstract

Predicted language variation in the LIB.

	In-group	**Out-group**
socially desirable	abstract	concrete
socially undesirable	concrete	abstract

Predicted language variation in the LIB.

But LIB defines abstractness ad-hoc based on word-lists of predicates — all adjectives are more abstract than all verbs, etc. Social desirability is a vague notion as well.

Can we do better?

With specificity and affect as holistic measures, we can design a new hypothesis quadrant:

	in-group	**out-group**
positive affect	low specificity	high specificity
negative affect	high specificity	low specificity

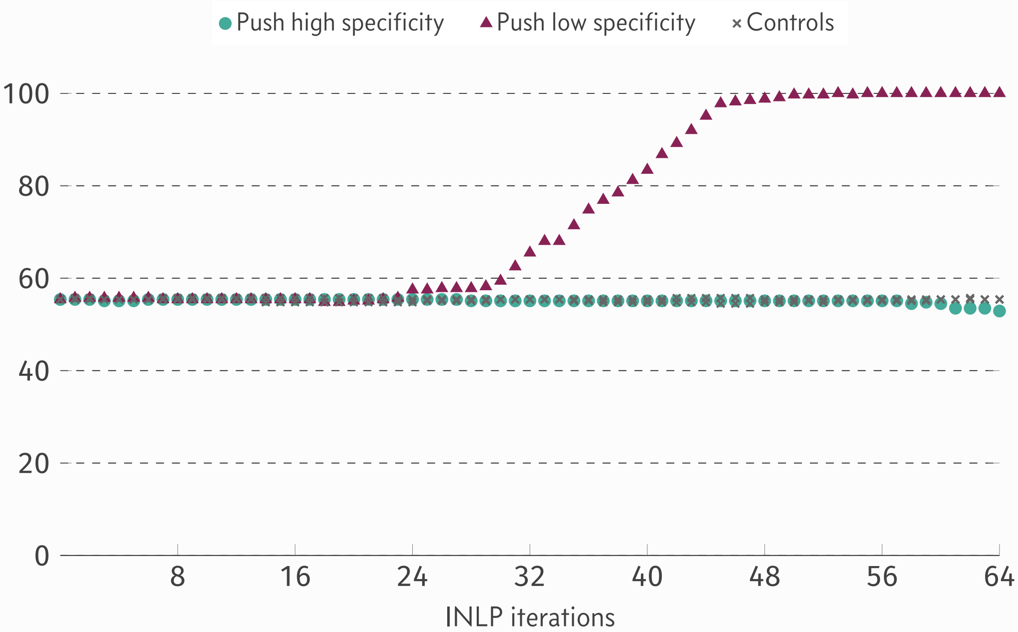
Predicted language variation in our more general formulation, using specificity and affect

With specificity and affect as holistic measures, we can design a new hypothesis quadrant:

	in-group	**out-group**
positive affect	low specificity	high specificity
negative affect	high specificity	low specificity

Predicted language variation in our more general formulation, using specificity and affect

We test this narrow hypothesis in Govindarajan et al., 2023, by **probing what our model learned**.


Push high specificity
Push low specificity
Controls
100
80
60
40
20
0
8
16
24
32
40
48
56
64
INLP iterations

- Our narrow, novel intergroup hypothesis didn’t replicate in the data, but that’s ok! It brings me back to the bigger picture.

- Our narrow, novel intergroup hypothesis didn’t replicate in the data, but that’s ok! It brings me back to the bigger picture.

- We need more natural language data to **discover** linguistic variations in how the intergroup bias is expressed.

- Our narrow, novel intergroup hypothesis didn’t replicate in the data, but that’s ok! It brings me back to the bigger picture.
- We need more natural language data to **discover** linguistic variations in how the intergroup bias is expressed.
- We need to account for the influence of real-world events which is the source of affect/emotion.

FOOTBALL FANDOM ON REDDIT

Previously, we were annotating tweets as a whole as in-group or out-group. But this doesn’t tell the whole story.

Previously, we were annotating tweets as a whole as in-group or out-group. But this doesn't tell the whole story.

③

a. **Mr. President-** Please tell your supporters to STAND DOWN, LEAVE the Capitol grounds and obey law enforcement who once again are risking their lives for our country!…

b. … Americans deserve answers on these unacceptable delays. We need a full accounting of **Pres Trump** and defense officials' decisions on Jan 6 …

c. We survived an insurrection and … In we made it clear what St. Louis already knew: **Donald Trump** was the white supremacist-in-chief.

① We need to look at references — does the intergroup bias influence how we **refer** to different groups?

① We need to look at references — does the intergroup bias influence how we **refer** to different groups?

② Probing suffered from utilizing two dimensions derived from the utterance — can we tie the utterance to a **non-linguistic description of events** preceding/precipitating the utterance?

① We need to look at references — does the intergroup bias influence how we **refer** to different groups?

② Probing suffered from utilizing two dimensions derived from the utterance — can we tie the utterance to a **non-linguistic description of events** preceding/precipitating the utterance?

③ How do we obtain language data with labelled intergroup information **at scale**?

- Reddit comments from **NFL game threads** on subreddits for each team.

- Reddit comments from **NFL game threads** on subreddits for each team.
- 568 games, 1104 threads, over 6 million comments.

- Reddit comments from **NFL game threads** on subreddits for each team.
- 568 games, 1104 threads, over 6 million comments.
- We have extensive documentation and statistics of every moment of the game, and reply on the NFLStats community for a simple, yet very effective **grounding** of utterances.

Win Probability (**WP**) is the probability of the in-group winning at a point in time. It is updated live with the game, with each play.

Win Probability (**WP**) is the probability of the in-group winning at a point in time. It is updated live with the game, with each play.

```
Win Probability (WP) = f(
                            seconds_remaining,
                            yard_line,
                            score_differential,
                            down,
                            Vegas_line,
                            ...
                         )
```

Eagles fans comments	Chiefs WP	Chiefs fans comments
Oh, is there a defense on the field?	0.75	Burn that clock baby

Comments from the Chiefs and Eagles fans with the WP for the Chiefs in the middle. The WP for the Eagles is 1-WP for the Chiefs.

4

a. Rams are gifting us a chance to win and we can’t take advantage. The fuck!!!!

b. if the ravens and chiefs beat these dudes by double digits then damn it so should we!

④ a. Rams are gifting us a chance to win and we can't take advantage. The fuck!!!!

b. if the ravens and chiefs beat these dudes by double digits then damn it so should we!

Why can't we treat this as a tagging/labelling task?

⑤ a. [OUT] are gifting [IN] a chance to win and [IN] can't take advantage. The fuck!!!!

b. if [OTHER] and [OTHER] beat [OUT] by double digits then damn it so should [IN]!

I annotated 1500 (random) comments for in-group, out-group (and other) by selecting spans from comments that correspond to these different groups. In addition to the comment, I was given the **live score**, source subreddit, and opponent.

I annotated 1500 (random) comments for in-group, out-group (and other) by selecting spans from comments that correspond to these different groups. In addition to the comment, I was given the **live score**, source subreddit, and opponent.

- 1500 comments from 768 threads and 491 games.

I annotated 1500 (random) comments for in-group, out-group (and other) by selecting spans from comments that correspond to these different groups. In addition to the comment, I was given the **live score**, source subreddit, and opponent.

- 1500 comments from 768 threads and 491 games.
- 1393 [IN] tags, 266 [OUT] tags, 166 [OTHER] tags.

I annotated 1500 (random) comments for in-group, out-group (and other) by selecting spans from comments that correspond to these different groups. In addition to the comment, I was given the **live score**, source subreddit, and opponent.

- 1500 comments from 768 threads and 491 games.
- 1393 [IN] tags, 266 [OUT] tags, 166 [OTHER] tags.
- 399 comments with *no annotation*.

⑥ a. Our oline should start holding since apparently it ’s okay now . Maybe Wilson can actually get some time to throw .

(6)

a. Our oline should start holding since apparently it ’s okay now . Maybe Wilson can actually get some time to throw .

b. [SENT] Lets go to the 4th with a 1st down around midfield.

(6) a. Our oline should start holding since apparently it 's okay now . Maybe Wilson can actually get some time to throw .

b. [SENT] Lets go to the 4th with a 1st down around midfield.

c. turning the game off , have a good day yall

(6)

a. Our oline should start holding since apparently it ’s okay now . Maybe Wilson can actually get some time to throw .

b. [SENT] Lets go to the 4th with a 1st down around midfield.

c. turning the game off , have a good day yall

People talk about the game, and refer to in-group/out-group in different ways. Does this change systematically with WP?

1. **Names of people**: *Tua, TK87, he/him,…*

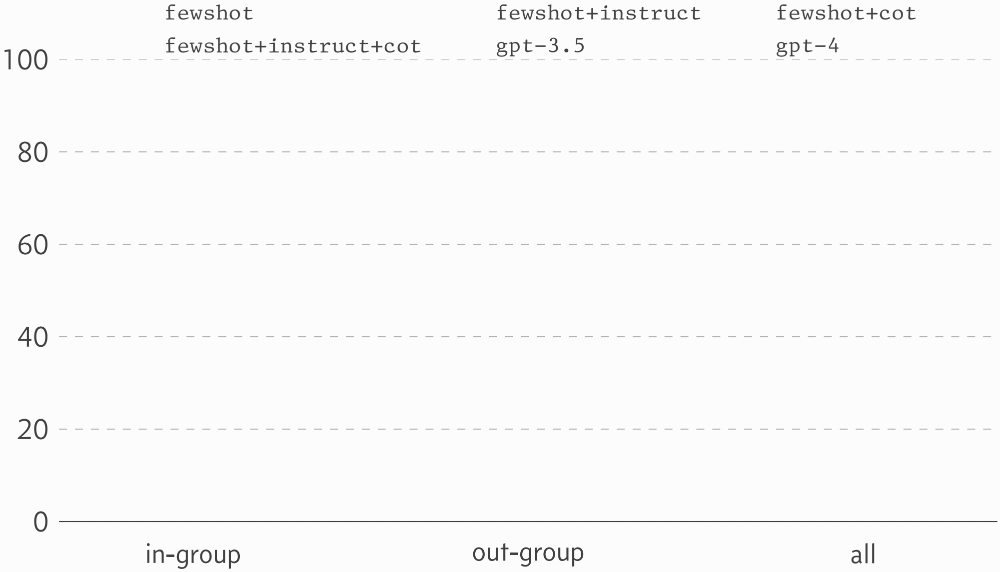
2. **Subset of the team**: *the offense*, *our defense*, *o-line*, …

3. **Team**: names (*rams, bills, cowboys*), nicknames (*lambs, cowgirls*), city names(*LA, Buffalo, Dallas*), pronominal expressions like *our boys*, pronouns like *they/them* for the in-group and out-group…

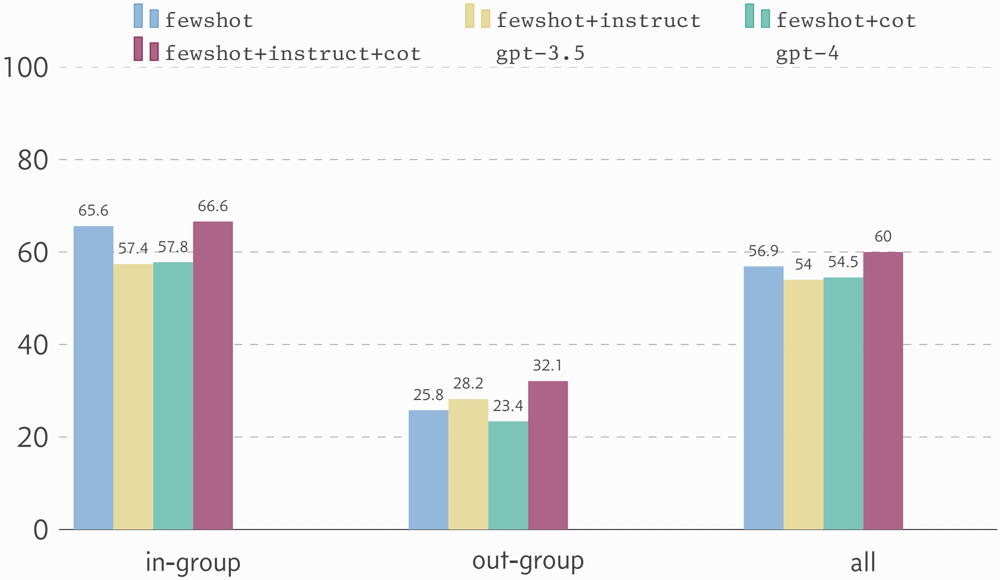
4. **Team plus supporters**: *we*, *us*, *they* and *them*

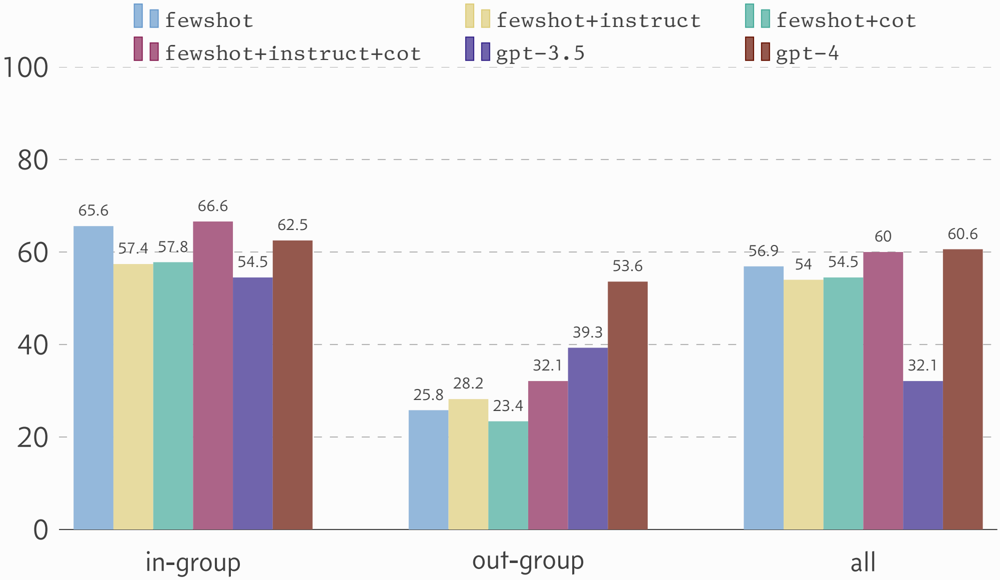
We want a large dataset of labelled/tagged comments, which is prohibitive with human annotation. Let's finetune an LLM with **instructions**, chain-of-thought **explanations** and **fewshot-examples**.

RESULTS-RECALL



RESULTS-RECALL





Instructions + examples + explanations works best.

Never bet against size (or GPT). I’m currently working towards finetuning Mistral.

Never bet against size (or GPT). I'm currently working towards finetuning Mistral.

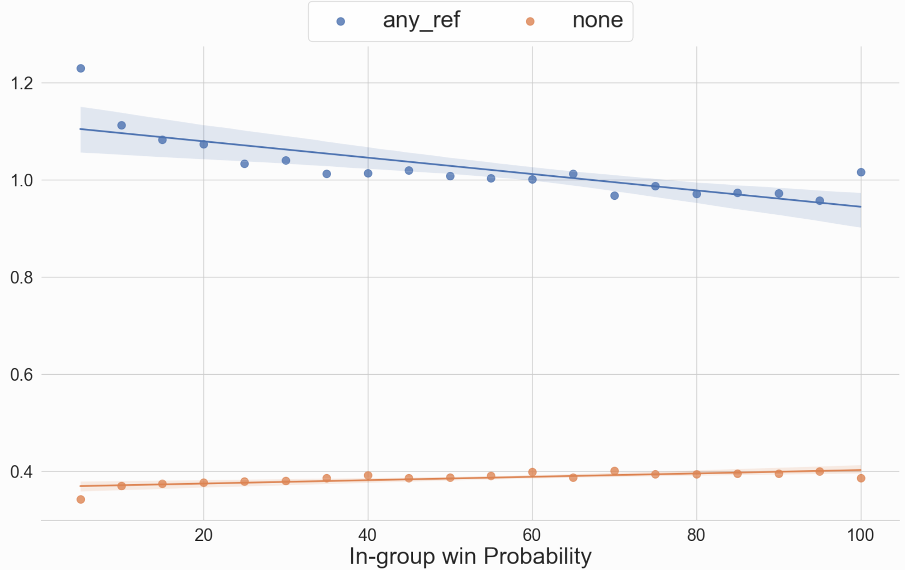
Modeling accuracy is **independent of WP** — most of the low performance on out-group is low recall. We can make inferences on a representative sample of comments labelled with in-group and out-group tags.

Never bet against size (or GPT). I'm currently working towards finetuning Mistral.

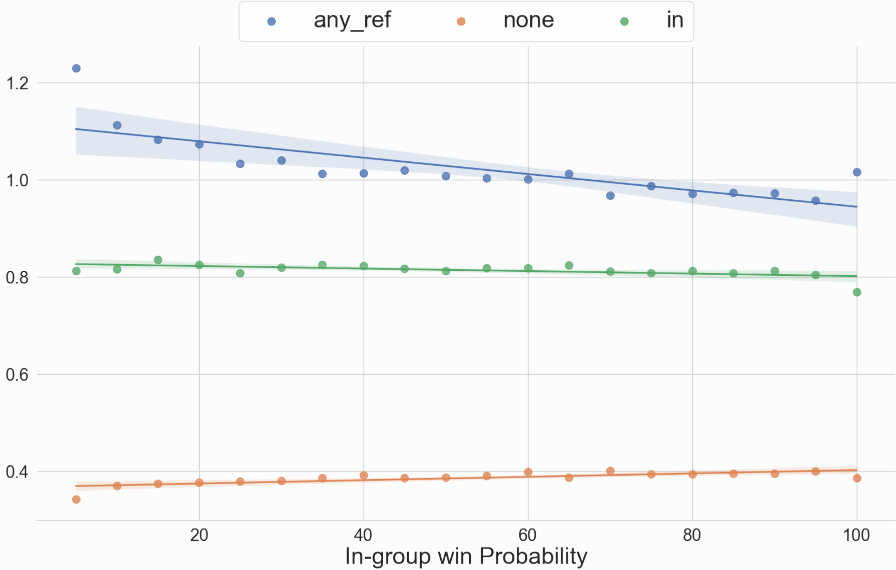
Modeling accuracy is **independent of WP** — most of the low performance on out-group is low recall. We can make inferences on a representative sample of comments labelled with in-group and out-group tags.

In analysis presented from now, I tagged over 200,000 randomly sampled comments using `fewshot+instruct+cot` model.

ALL GOES DOWN



Frequency of any-group and null references over all 5% WP windows from 0 to 100



Frequency of any-group, null and in-group (normalized within any-group) references over all 5% WP windows from 0 to 100

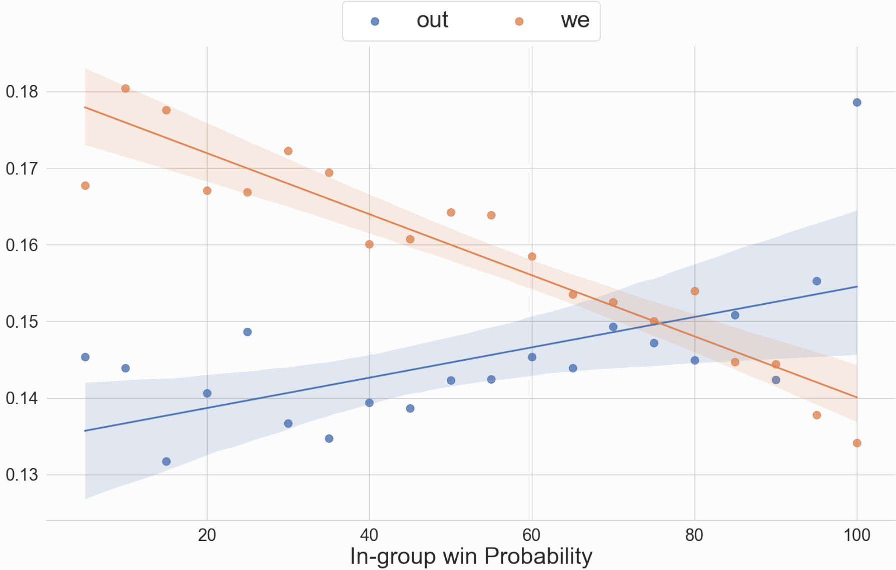
The better the state of affairs in the real world for the *in-group*, the more likely commenters are to **abstract away** from specifically referring to the in-group (or any group).

The better the state of affairs in the real world for the *in-group*, the more likely commenters are to **abstract away** from specifically referring to the in-group (or any group).

(7)
a. HOLY SHIT

b. DO NOT TAKE YOUR FOOT OFF THE GAS

c. WHAT A THROW



Frequency of references to the in-group with first person plural forms, and the out-group, over all 5% WP windows from 0 to 100.

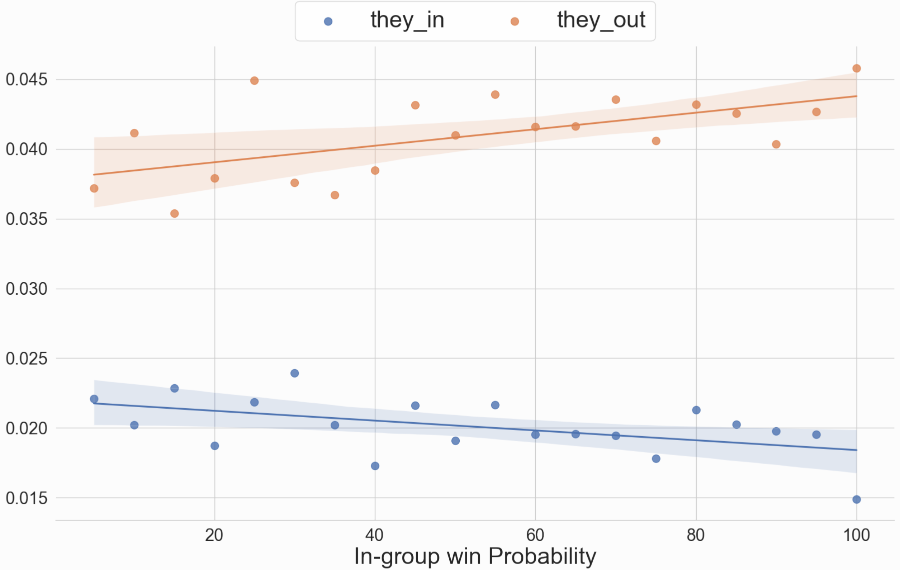
Fans band together to talk about the in-group with **third person plural referents** the less likely they are to win:

⑧ a. We need a reliable safety ..... like , BAD .

b. Our defense is not why we lost the game .

Fans prefer to talk (shit) about the out-group, and refer less to the in-group, when the in-group is doing well:

9
a. Keep going , hang 50 on these fuckers.
b. "Karma will get the Cowboys for trying to run up the score" Actual comment in the Falcons sub , tells you all you need to know about the pussies over there.



Frequency of references to the in-group or out-group with **third person plural** over all 5% WP windows from 0 to 100.

Intergroup bias Data curation and modeling can enable us to detect and study extremely subtle linguistic variations like intergroup bias at scale.

Intergroup bias Data curation and modeling can enable us to detect and study extremely subtle linguistic variations like intergroup bias at scale.

Grounding Win Probability is a well calibrated predictor of intergroup bias, as evidenced by the linear relationship in referential expression bias.

Intergroup bias Data curation and modeling can enable us to detect and study extremely subtle linguistic variations like intergroup bias at scale.

Grounding Win Probability is a well calibrated predictor of intergroup bias, as evidenced by the linear relationship in referential expression bias.

Future Explore the **parallel** nature of the corpus further, **multilingual** work, circling back to **stereotypes**...

REFERENCES

Van Dijk, Teun A (2009). *Society and Discourse: How Social Contexts Influence Text and Talk*. Cambridge University Press.

Beaver, David and Jason Stanley (2018). "Toward a Non-Ideal Philosophy of Language". In: *Graduate Faculty Philosophy Journal* 39.2, pp. 503–547.

Kaneko, Masahiro and Danushka Bollegala (July 2019). "Gender-preserving Debiasing for Pre-trained Word Embeddings". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1641–1650.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi (July 2020). "Social Bias Frames: Reasoning about Social and Power Implications of Language". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5477–5490.

Webson, Albert, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick (Nov. 2020). "Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4090–4105.

Maass, Anne (Jan. 1, 1999). “Linguistic Intergroup Bias: Stereotype Perpetuation Through Language”. In: *Advances in Experimental Social Psychology*. Ed. by Mark P. Zanna. Vol. 31. Academic Press, pp. 79–121.

Gorham, Bradley W. (2006). “News Media’s Relationship With Stereotyping: The Linguistic Intergroup Bias in Response to Crime News”. In: *Journal of Communication* 56.2. Place: United Kingdom Publisher: Blackwell Publishing, pp. 289–308. issn: 1460-2466(Electronic),0021-9916(Print).

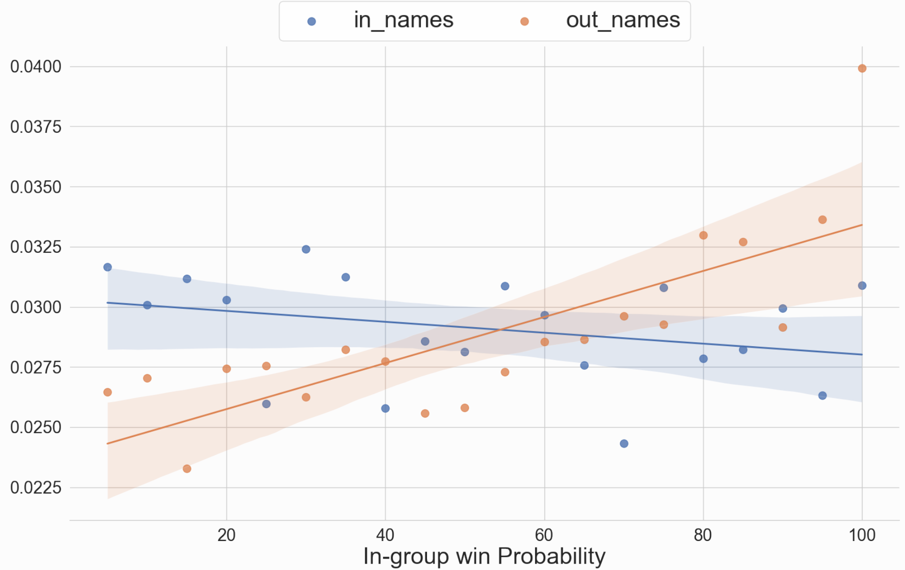
Sainburg, Tim, Leland McInnes, and Timothy Q Gentner (2021). “Parametric UMAP Embeddings for Representation and Semisupervised Learning”. In: *Neural Computation* 33.11, pp. 2881–2907.

Govindarajan, Venkata S, David Beaver, Kyle Mahowald, and Junyi Jessy Li (July 2023). “Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 12853–12862.

FEWSHOT EXAMPLE

```
COMMENT: [SENT] Defense getting absolutely bullied by a dude that looks
         like he sells solar panels
IN-GROUP: Jets
OUT-GROUP: Bears
WIN PROBABILITY: 71.5%
TARGET: [SENT] [IN] getting absolutely bullied by [OUT] that looks
        like [OUT] sells solar panels .
REF_EXPRESSIONS: ['Defense', 'a dude', 'he']
```

NAMES



Frequency of references to the in-group or out-group by name over all 5% WP windows from 0 to 100.

Feature	Slope	r-squared
Any reference	-17×10^{-4}	0.61
No reference	3.5×10^{-4}	0.57

SLOPES OF SIGNIFICANCE

Feature	Slope	r-squared
Any reference	-17×10^{-4}	0.61
No reference	3.5×10^{-4}	0.57
In-group	-2.6×10^{-4}	0.31
we/us	-4×10^{-4}	0.87
they_in	-4×10^{-5}	0.22
they_out	6×10^{-5}	0.33

Table of slopes of feature of interest against increasing WP, alongside the r-squared showing how much of the variance is explained by the linear regression fit.