Counterfactual Probing for the Influence of Affect & Specificity on Intergroup Bias

Venkat, David Beaver, Kyle Mahowald, Jessy Li

venkatasg@utexas.edu

The University of Texas at Austin

In Govindarajan (2023) @ EACL 2023, we introduced framing bias in terms of the **intergroup relationships** between the speaker and the target of their utterance.

All language is biased. Intergroup relationships are a *social influence on language production*, and we aim to directly model it.

In Govindarajan (2023) @ EACL 2023, we introduced framing bias in terms of the **intergroup relationships** between the speaker and the target of their utterance.

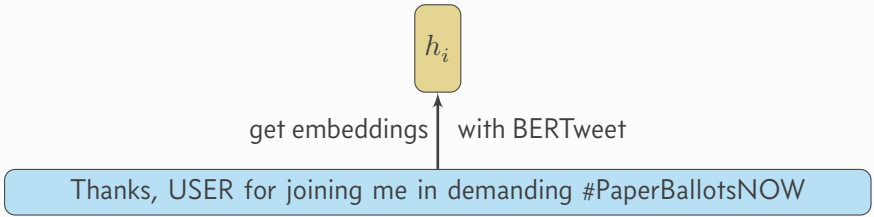
All language is biased. Intergroup relationships are a *social influence on language production*, and we aim to directly model it.

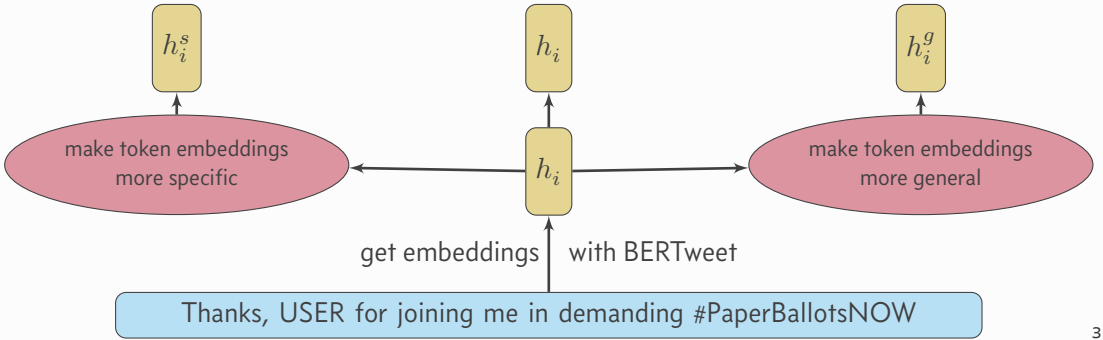
Which linguistic features change systematically in **in-group** vs. **out-group** contexts?

Affect is a coarse grained feature that estimates how a speaker feels towards the target they mentioned in an interpersonal utterance.

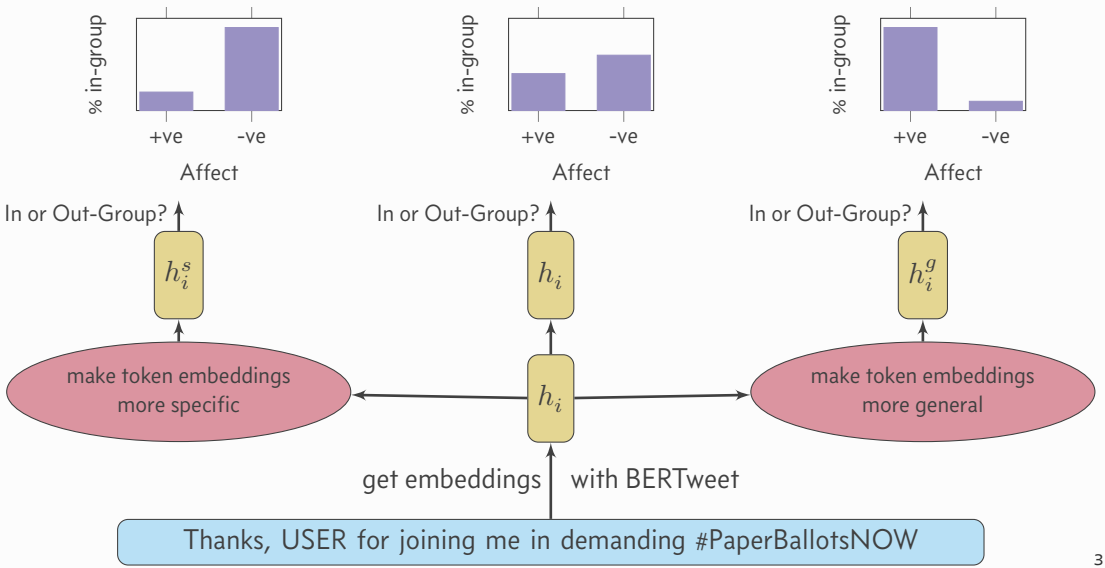
Specificity measures the level of detail and involvement of concepts, objects and events.

Thanks, USER for joining me in demanding #PaperBallotsNOW

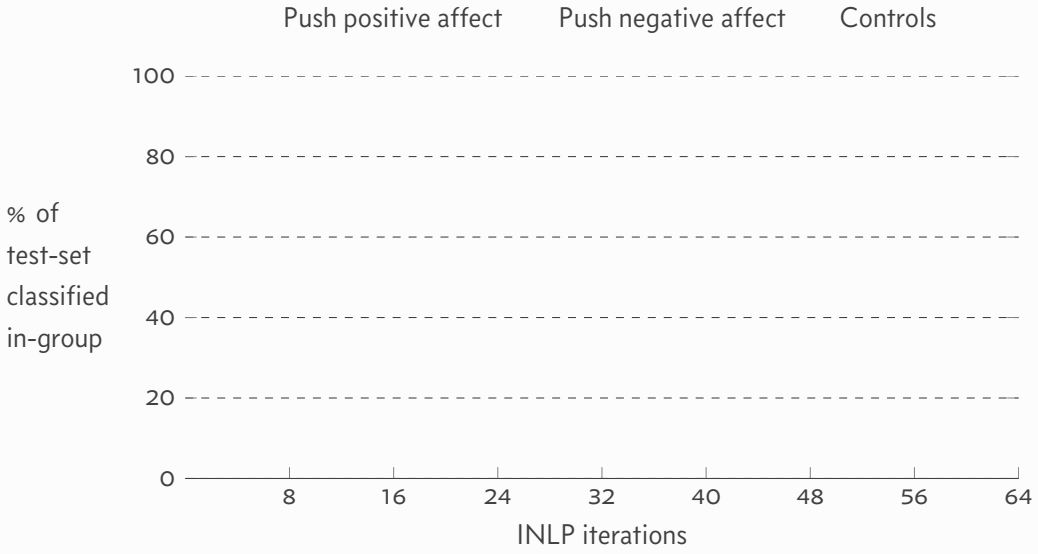

h_i
get embeddings
with BERTweet
Thanks, USER for joining me in demanding #PaperBallotsNOW


h_i^s
h_i
h_i^g
make token embeddings more specific
h_i
make token embeddings more general
get embeddings
with BERTweet
Thanks, USER for joining me in demanding #PaperBallotsNOW

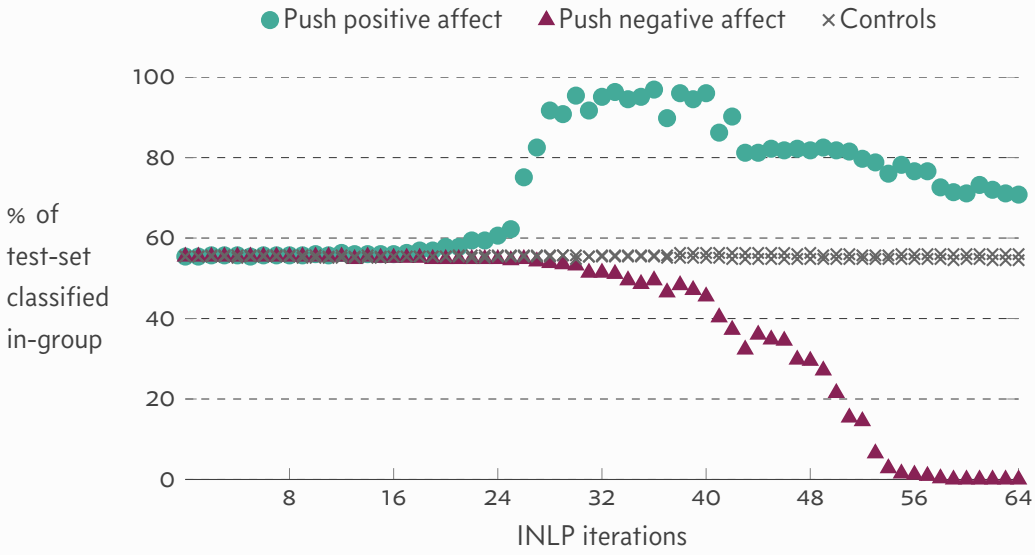
PROBING HYPOTHESIS



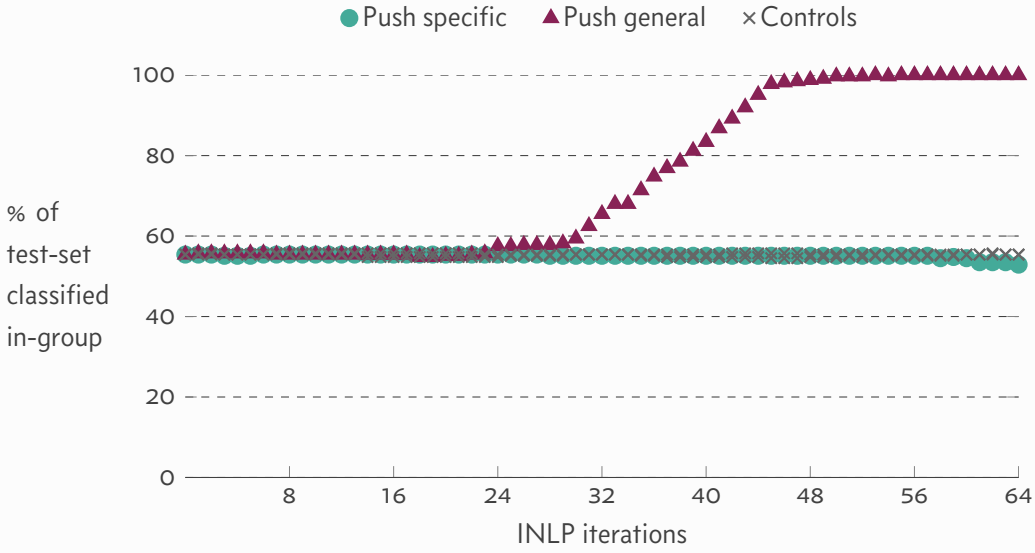
AFFECT RESULTS



AFFECT RESULTS



SPECIFICITY RESULTS



Come by our poster!

Check our our paper online!