

Computational Models of Social Meaning

Venkata S Govindarajan

Feb 2 2024

The University of Texas at Austin

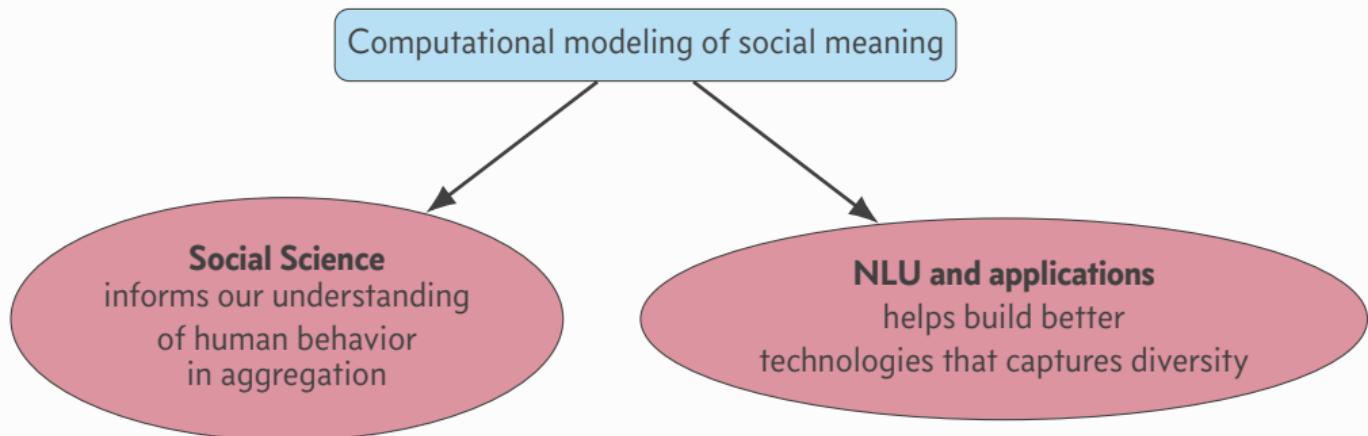
How does our **social identity** interact with our linguistic behavior?

...the constellation of qualities and properties that linguistic forms convey about the social identity of language users

Beltrama, 2020

All linguistic behavior is **situated in social context**, is informed and moulded by it, and actively constructs it as well.

WHY MODEL SOCIAL MEANING?



AN ILLUSTRATING EXAMPLE

- ① 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 are **subtle cues of social identity** (in-group in (a) and out-group in (b)) .

Computational modeling on data can reveal these at scale.

SIMPLE MODELS CAN REVEAL BASIC INTUITIONS

In-group	Out-group
thanks, love, count me	thanks, bipartisan, restore
birthday, my colleague	kind, resignation

Top unigram and bigram features from a Naive Bayes model for in-group vs. out-group on our dataset.

We will discuss this and further results from "[How people talk about each other: Modeling Generalized Intergroup Bias and Emotion](#)"(EACL 2023)

ANOTHER ILLUSTRATION

- ② a. I **would encourage** finding a therapist to add to your support team.
- b. I **talked** on Reddit with others to get support and ideas .

Both utterances are responses to an online post asking for advice — both of them **give advice**, the latter is more implicit.

Computational models can uncover diverse ways of doing things with language, and their behavior can be informative. We will touch on results from "[Help! Need Advice on Identifying Advice](#)"(EMNLP 2020).

A FINAL EXAMPLE PAIR

- ③ a. Hurts really **overcame** some **adversity** to **fumble** on that 3rd down .
- b. Well **that sucked**.

These are comments **describing the same event** from different social perspectives (out-group in (a) and in-group in (b)).

Computational modeling on data can discern how we abstract/generalize over people and events, and the myriad ways in which we do it.

Diversity/Variation

What are the social pressures influencing the diversity in form of language, and what is the nature of their influence?

Generalization

How do we generalize over events and entities to communicate socially and contextually relevant meaning?.

① **How does our social identity shape linguistic actions?**

"Help! Need Advice on Identifying Advice"(EMNLP 2020)

② **How does intergroup social identity lead to implicit bias in our language?**

"How people talk about each other: Modeling Generalized Intergroup Bias and Emotion"(EACL 2023)

"Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias"(Findings@ACL 2023)

③ **Ongoing work grounding the intergroup bias, and a peek into my future research plans.**

① **How does our social identity shape linguistic actions?**

"Help! Need Advice on Identifying Advice"(EMNLP 2020)

② **How does intergroup social identity lead to implicit bias in our language?**

"How people talk about each other: Modeling Generalized Intergroup Bias and Emotion"(EACL 2023)

"Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias"(Findings@ACL 2023)

③ **Ongoing work grounding the intergroup bias, and a peek into my future research plans.**

One of the hallmarks of human language is that we use it to **do things** in the world.

- ④ a. I do (sc. take this woman to be my lawfully wedded wife).
- b. I name this ship Boaty Mcboatface.
- c. I suggest you try a calming spray.

How do we **use our social identities and awareness** to effectively do things?

In Govindarajan, Chen, et al., EMNLP 2020, we investigated one form of performative language use: **giving advice online**.

How is advice **structured** online?

How good are computational models at **identifying advice**?

Parenting with a history of depression?

- ⑤ **I took** my meds the whole time. **I used** the tools I learned in therapy. **I talked** on Reddit with others to get support and ideas.

(r/AskParents)

People often give advice **implicitly** using personal narratives and other strategies (Abolfathiasl et al., 2013).

To model general online human advice-seeking interactions, we chose to construct a dataset from 2 English Reddit forums (subreddits) focused on advice.

r/AskParents	r/needadvice
parents seeking advice	a general advice forum

[GITHUB.COM/VENKATASG/ADVICE-EMNLP2020](https://github.com/venkatasg/advice-emnlp2020)

16.4% of advice sentences from r/AskParents, but only **7%** from r/needadvice were judged to contain **personal narratives** from a sample analysis.

How well do models identify advice, especially advice from personal narratives?

We model advice identification as a **binary classification task**.

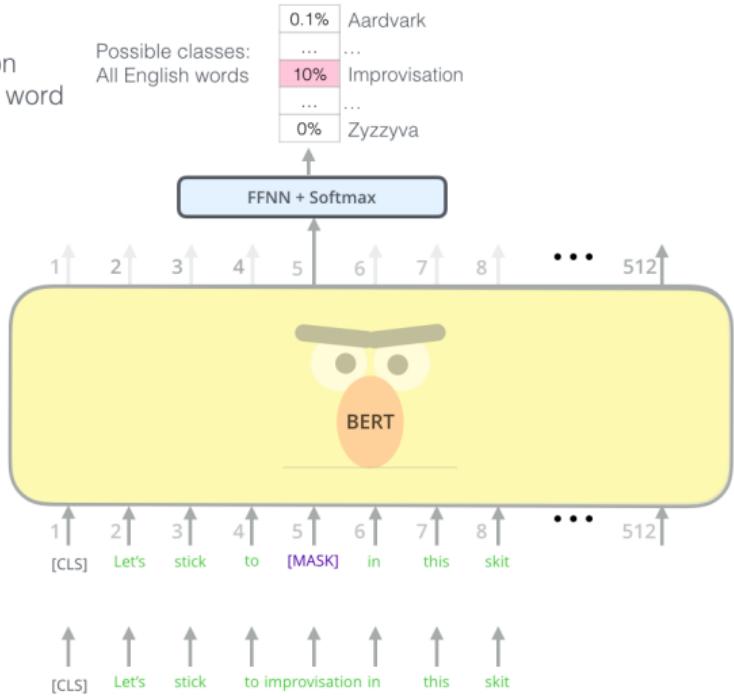
Rule-based Match and score against **words, phrases, regexs**:

*suggest, recommend, . *would\slike. *if. **

Learned embeddings We use **BERT** and perform binary classification on a sentence representation after further training on our dataset.

WHAT IS BERT?

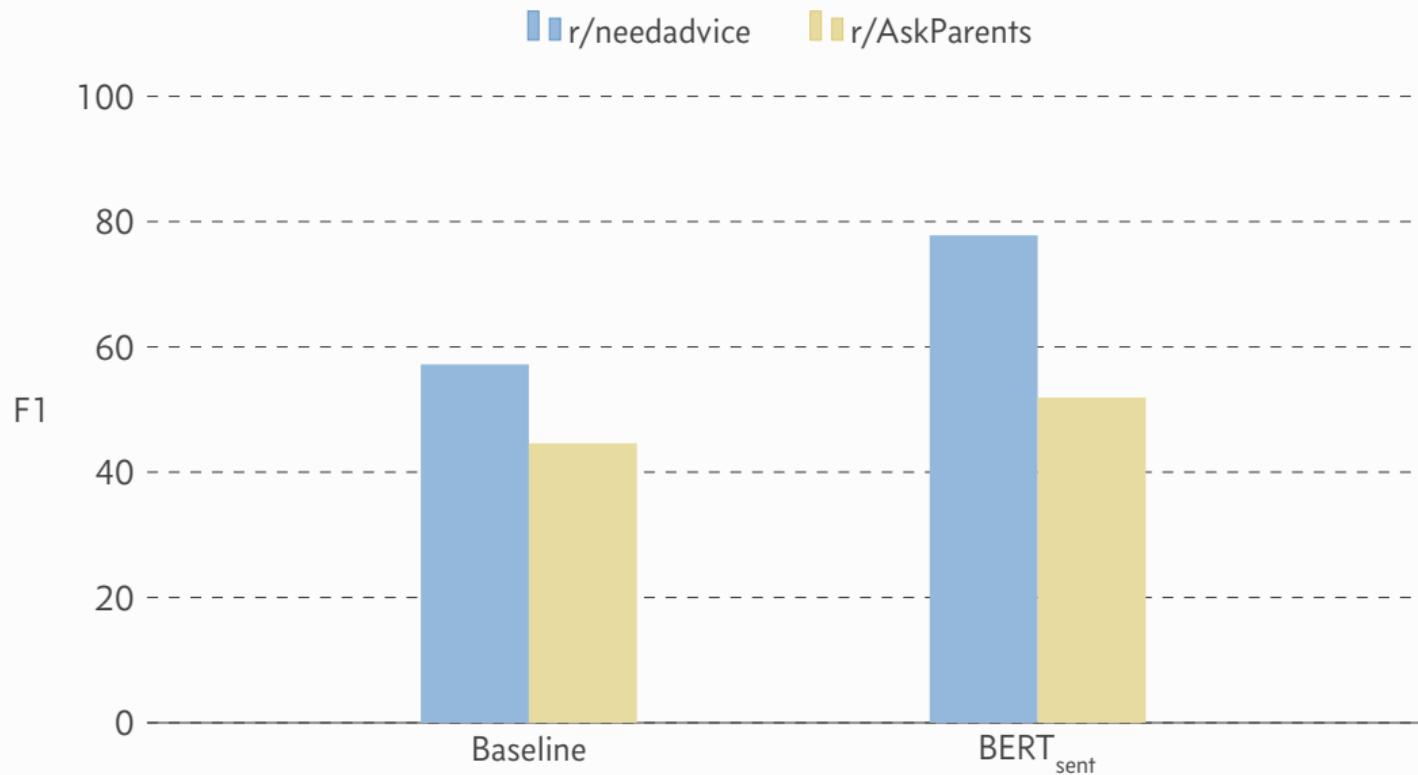
Use the output of the masked word's position to predict the masked word



Randomly mask 15% of tokens

Through pre-training on word prediction on large corpora, neural language models compress and **encode** a large amount of linguistic information. Illustration from *The Illustrated BERT, ELMo, and co. (Blog post)* (2020)

RESULTS - IDENTIFYING ADVICE



Performance on r/AskParents worse than r/needadvice. Why? Personal narratives.

Personal narratives Dataset shows the importance of how personal narratives shaped by a shared social identity inform effective advice-giving.

Model weaknesses The best models perform worse at identifying implicit advice given through personal narratives — its understanding of what constitutes advice is relatively surface-level.

Advice-giving is (mostly) **intentional**, and people (naturally) use their social identity and knowledge of social relationships to give effective advice.

But this is not always be the case — do aspects of our social identity and relationships impact our language use **unconsciously**?

① **How does our social identity shape linguistic actions?**

"Help! Need Advice on Identifying Advice"(EMNLP 2020)

② **How does intergroup social identity lead to implicit bias in our language?**

"How people talk about each other: Modeling Generalized Intergroup Bias and Emotion"(EACL 2023)

"Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias"(Findings@ACL 2023)

③ **Ongoing work grounding the intergroup bias, and a peek into my future research plans.**

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.

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

- ⑥ 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.

- ⑦ 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.

Analyze and model 2 dimensions of intergroup bias — **intergroup relationship** and **interpersonal emotion**.

How does intergroup relationship (in-group vs. out-group) **interact** with interpersonal emotion?

We answer these questions in [Govindarajan, Atwell, et al., EACL 2023](#), and start with some clear definitions:

Interpersonal Utterance is any utterance where there is a target individual talked about or referred to.

Intergroup Relationship is defined as the relationship between the speaker and target of an utterance — in-group or out-group.

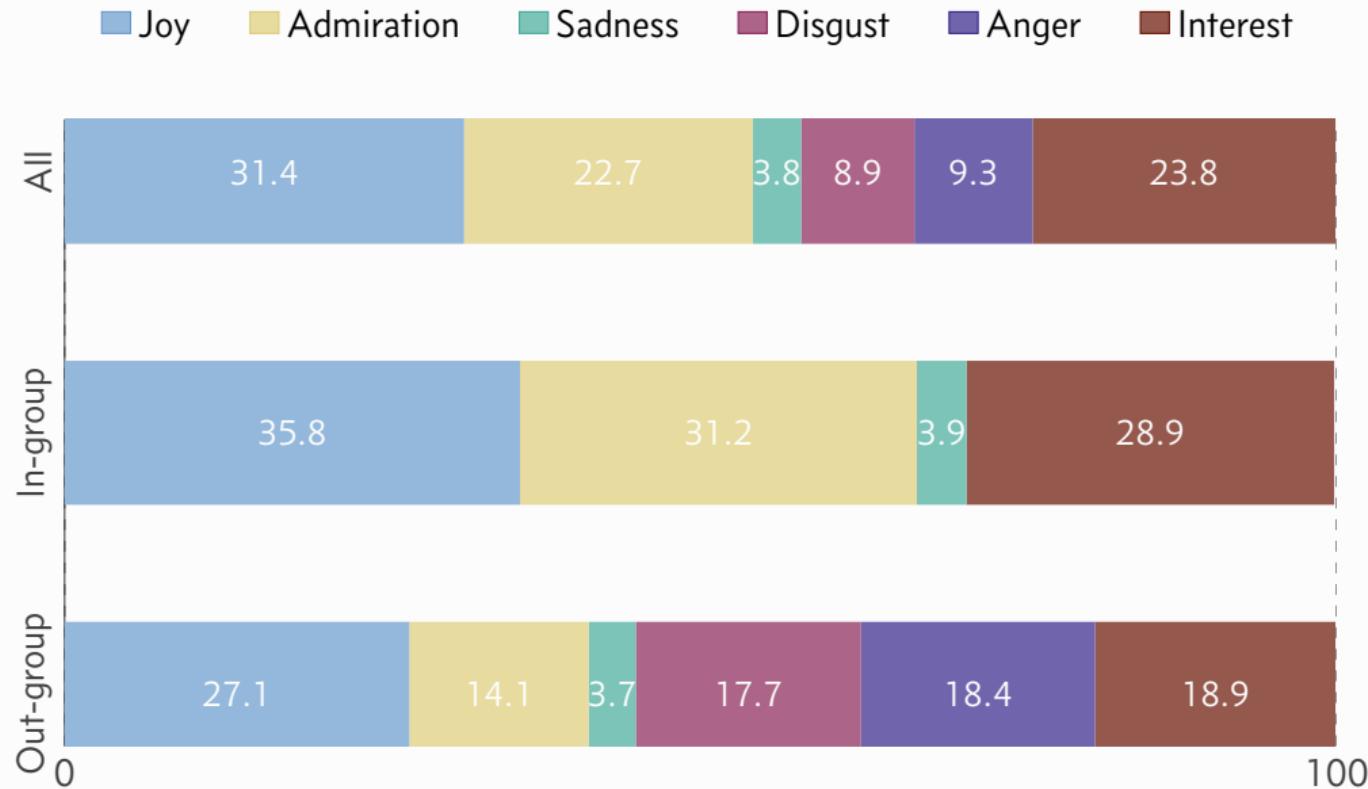
Interpersonal Emotion is defined as the emotion expressed by a speaker s towards, or in connection with the target t of the utterance u, as perceived by a reader.

- Tweets by members of US Congress which mention one other member.
- Tweets are either directed in-group or out-group.

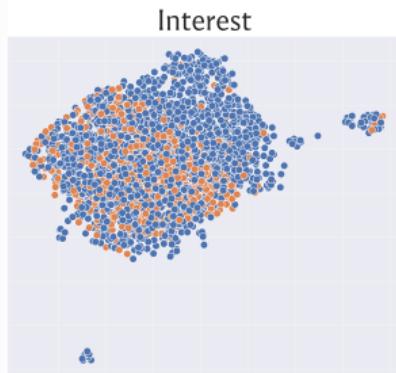
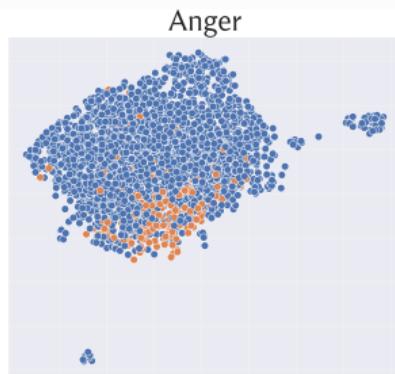
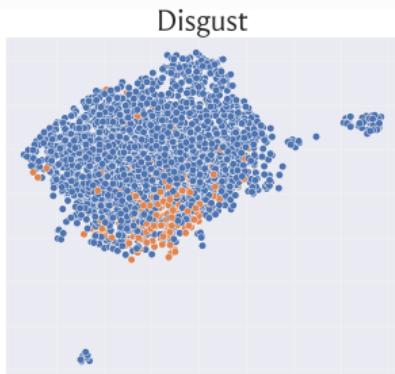
3033 (identity masked) tweets annotated for fine-grained emotion using Plutchik wheel, with *found supervision* for intergroup relationship labels.

[GITHUB.COM/VENKATASG/INTERPERSONAL-BIAS](https://github.com/venkatasg/INTERPERSONAL-BIAS)

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.

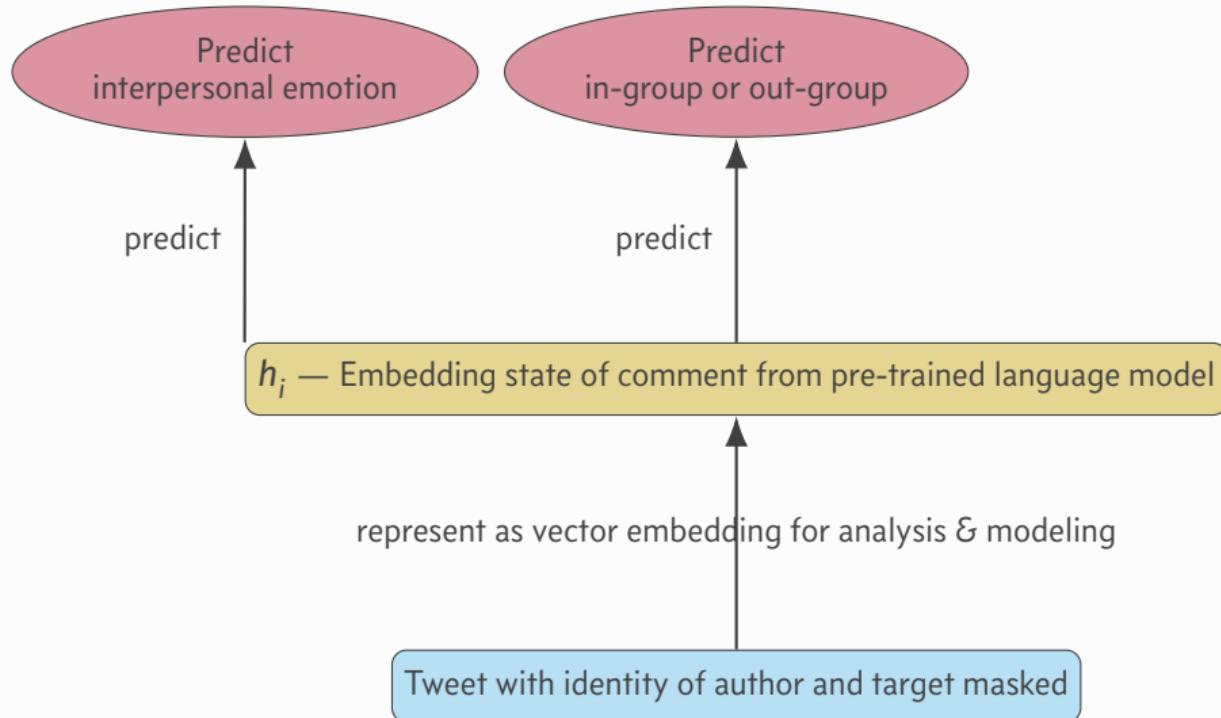
Two tasks: predict **Intergroup Relationship** and **Interpersonal Emotion**.

Baseline Predict Intergroup Relationship with NB-SVM with unigrams and bigrams, and Interpersonal Emotion with EMOLEX.

BERTweet Predict both dimensions with classification or labelling layer on top of finetuned BERTweet embeddings.

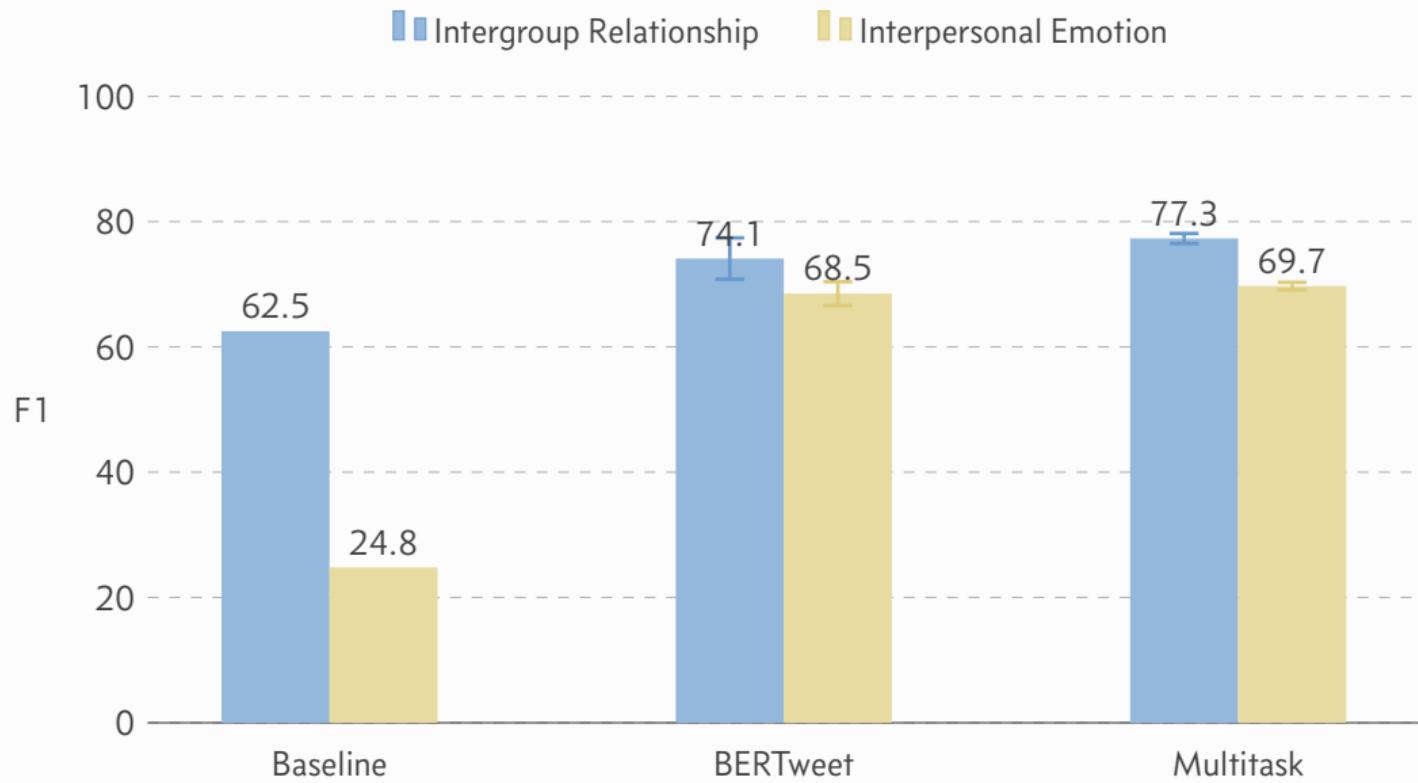
Multitask Predict both dimensions simultaneously with shared BERTweet encoding to see if they mutually support each other.

WHAT IS MULTITASKING?



Train model end-to-end so that emotion **transfers** information for in/out disambiguation and vice-versa.

RESULTS



Multitasking improves on vanilla model (slightly).

- 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**?

The Linguistic intergroup bias hypothesized that stereotypes persist by systematic differences in how people **generalized** over events depending on the target individual in an utterance.

- ⑧ 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**.

	Socially desirable	Socially undesirable
in-group	abstract	concrete
out-group	concrete	abstract

Predicted language variation in 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?

Specificity is a pragmatic concept of text that measures the **level of detail** and involvement of concepts, objects and events.

- ⑨ a. Couldn't agree more @Doe. **GENERAL**
- b. I'm trying to shift the focus to smaller businesses and workers so we can get money directly to the American people. cc: @Doe. **SPECIFIC**
- c. It was wonderful to celebrate the 150th anniversary of Washington's Pilotage Act this week with @USER and this fantastic group in Port Townsend **VERY SPECIFIC**

Using lexical and sentence-level hand-curated and learned features, we can **automatically assign specificity ratings** (1-5) to tweets in our dataset.

Affect measures how the speaker **regards** the target of their interpersonal utterance:

- (10) a. @USER was a true patriot...I am sad ...**POSITIVE**
- b. We can't let @USER succeed in silencing a war debate that Congress should have held years ago. **NEGATIVE**

We **annotate for affect** on our Congressional tweets dataset — annotation and preprocessing derives a binary affect label.

NEW (NARROW) HYPOTHESIS

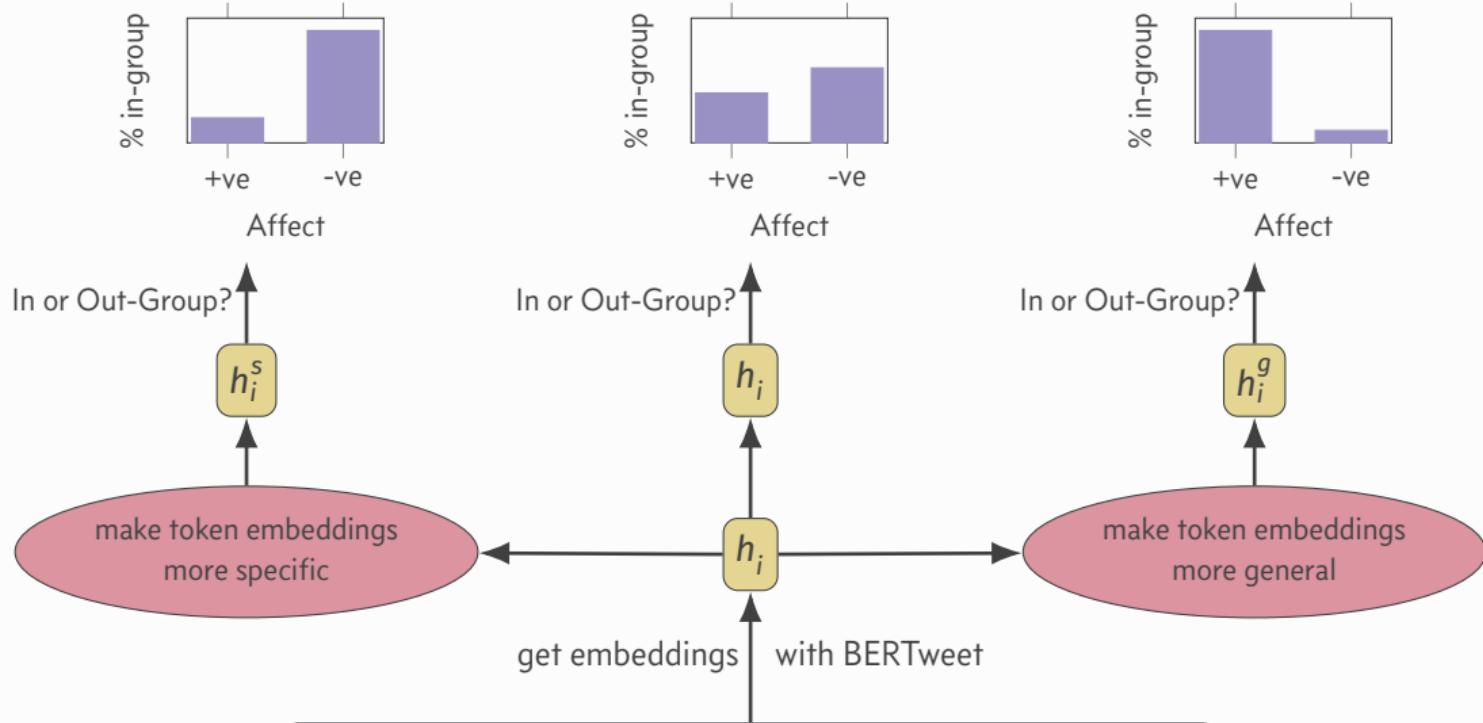
With specificity and affect as holistic measures, we can design a new hypothesis quadrant:

	positive affect	negative affect
in-group	general	specific
out-group	specific	general

Predicted language variation with new parameters.

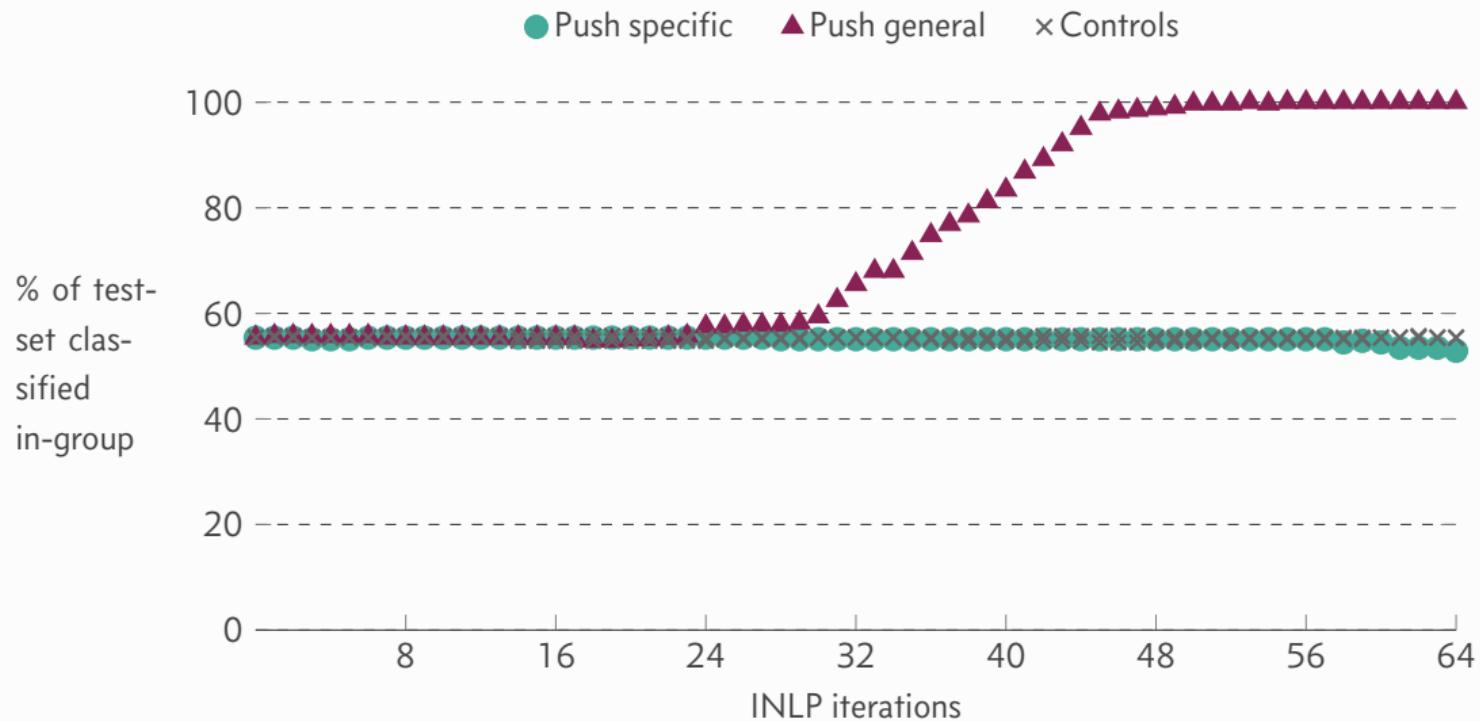
We test this narrow hypothesis in [Govindarajan, Beaver, et al., Findings@ACL 2023](#), by **prob-ing what our model learned**.

PROBING HYPOTHESIS



Thanks, @USER for joining me in demanding #PaperBallotsNOW

SPECIFICITY RESULTS



No difference in model predictions between positive and negative affect sentences in test set.

- 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 in addition to affect/emotion.

① **How does our social identity shape linguistic actions?**

"Help! Need Advice on Identifying Advice"(EMNLP 2020)

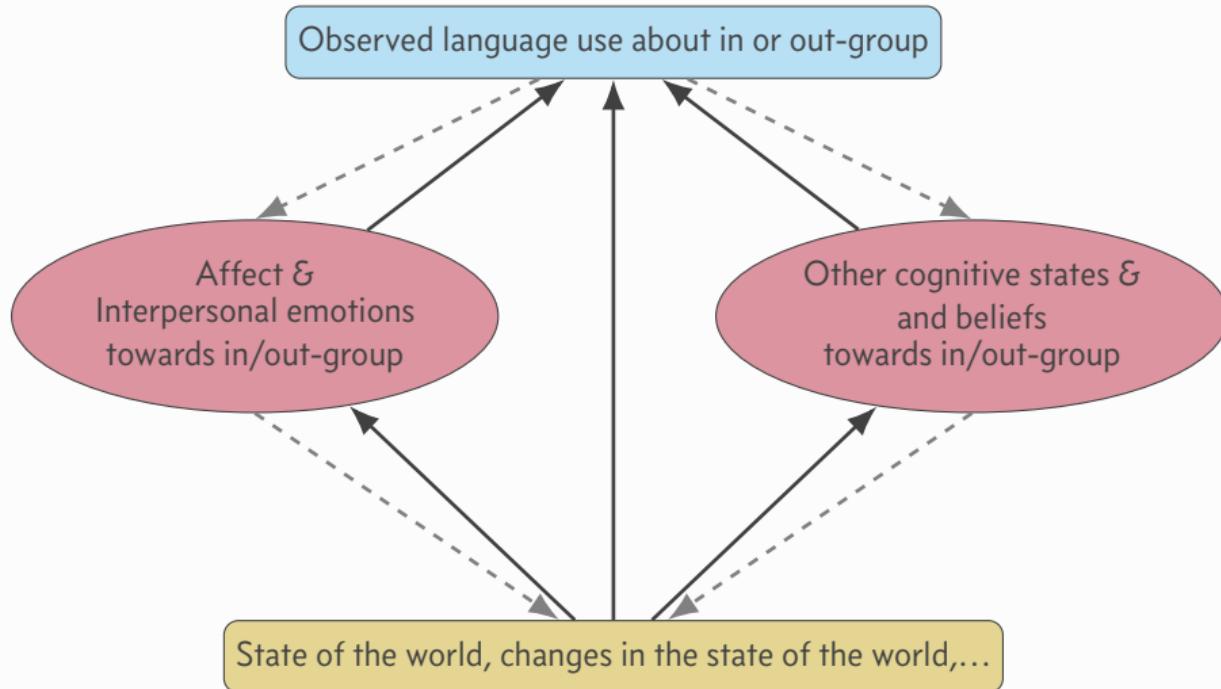
② **How does intergroup social identity lead to implicit bias in our language?**

"How people talk about each other: Modeling Generalized Intergroup Bias and Emotion"(EACL 2023)

"Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias"(Findings@ACL 2023)

③ **Ongoing work grounding the intergroup bias, and a peek into my future research plans.**

WHERE DOES THE BIAS COME FROM?



Language affects the cognitive state and world state too, but we're not getting into that now.

GROUNDING THE INTERGROUP BIAS

We want a new dataset of natural language use, where we can compare and contrast in-group and out-group speech from both perspectives, and ground it in real-world events.

Posted by u/nfl_gdt_bot 3 days ago
308 Game Thread: Kansas City Chiefs (11-6) at Buffalo Bills (11-6)

Kansas City Chiefs at Buffalo Bills
[ESPN Gamecast](#)
Highmark Stadium- Orchard Park, NY
Network(s) Waterwagon_78 · 2 days ago
So can the naysayers finally agree that this is a dynasty? 3 superbowl appearance, 2 SB WINS!, 6 straight AFC championship appearances!
3 1 Reply Share ...

TacoGuardsman · 2 days ago
For sure is
1 Reply Share

Posted by u/AutoModerator 3 days ago 🔒
461 [GDT] Buffalo Bills vs. Kansas City Chiefs
Gameday

- Please be mindful of the rules
- Please report any violations
- Self-posts will subject to deletion
- Go Bills!

17.8k Comments Share Save ...

FreePop5311 · 3 days ago
[Disney Pixar via GIPHY](#)

by Disney Pixar via GIPHY
78 Share ...

2kptr · 3 days ago
CUT THIS SHIT ASS KICKER TOMORROW
118 Share ...

About Community
r/KansasCityChiefs
Home of the Kansas City Chiefs Subreddit
Created Sep 3, 2010

About Community
r/buffalobills
The home of the Buffalo Bills on reddit.
Circling the wagons since 2009.
Created Oct 17, 2009

220k • 692 Top 1%
mathos right here, right now Ranked by Size

With Yuki Zang (undergraduate at Brown), I am analyzing a fresh source of rich expressions of in-group and out-group directed language: **reddit comments on NFL game threads**.

- Over 7 million comments for over 500 games, from **both perspectives**.
- Comments are time aligned with the official **play-by-play**, a non-linguistic description of events and its relevance to each group (effectively the scoreboard)

Why sports comments?

Diversity

What are the social pressures influencing the diversity in form of language, and what is the nature of their influence?

Generalization

How do we generalize over events and entities to communicate socially and contextually relevant meaning?.

Let me talk you through how they guide my approach to on-going work, and my long-term research program.

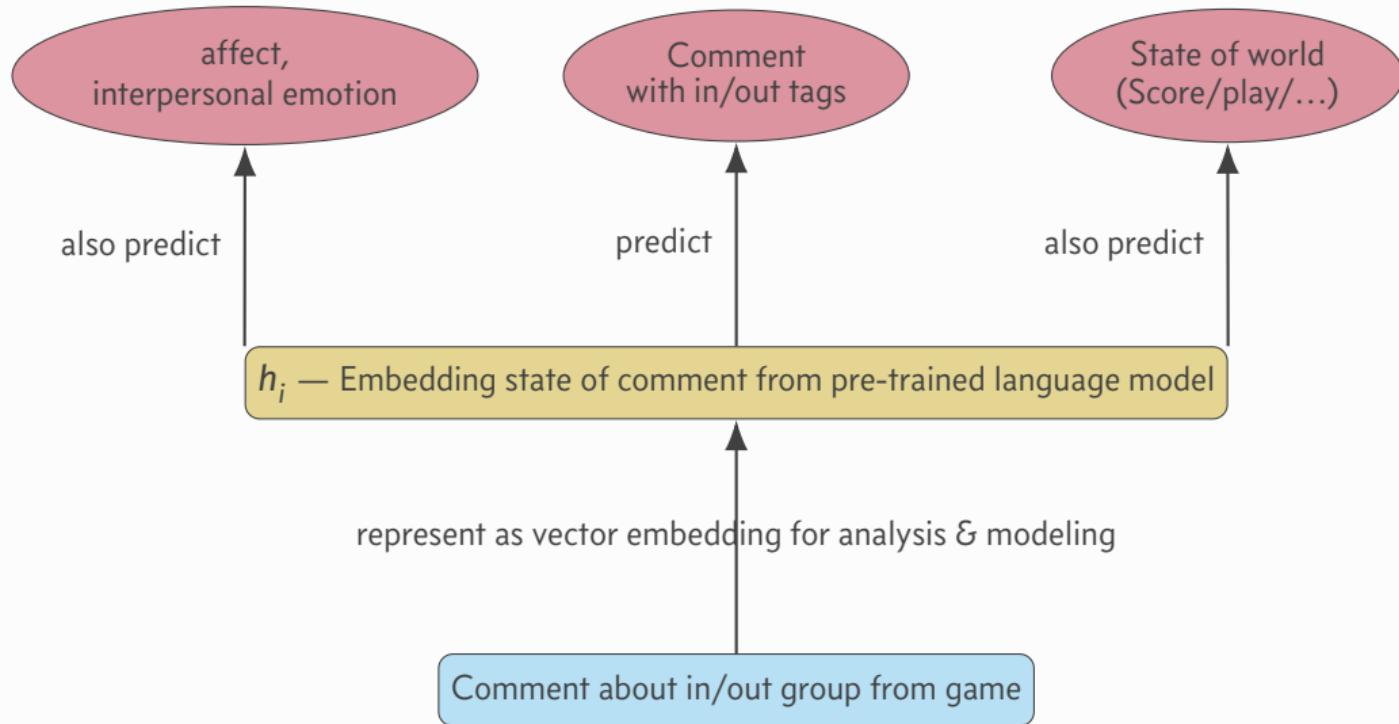
Chiefs comments	Play	Eagles comments
Did not expect their o-line to bully our defense like that	Touchdown—Eagles	That was some eagles football right there boys.
TK87 big money !!!	Touchdown—Chiefs	Is the defense back in Philly ?

Aligned in-group & out-group comments with events from Super Bowl 2023

- (11) a. Did not expect **their o-line** to bully **our defense** like that
 - b. That was some **eagles** football right there boys.
-
- (12) a. Did not expect **[OUT]** to bully **[IN]** like that
 - b. That was some **[IN]** football right there boys.

Through this reframing, we can approach detecting in-group vs. out-group speech as a **tagging** or **masked** prediction task — predict the tag (**[IN]** or **[OUT]**) based on context.

MULTITASKING



Train representation model end-to-end so that multi-tasking **transfers** information for in/out disambiguation.

Multitasking tests the model's ability to use emotion and world-state to identify diverse expressions of social meaning.

Short-term How do social identity, interpersonal emotion **and situational context** interact towards final linguistic form? What do we infer from model behavior?

Medium-term Diversity across dialects and languages with aligned events.

English commentator	Play	Spanish commentator
Gano from 63 yards out..it is good!	Field goal by Gano	GANO LO GANO! GANO LO GANO! GANO LO GANO!!

Long-term Social meaning diversity in non-polarized settings, intersectional identity.

Back to our examples:

- (13) a. Did not expect **their o-line** to bully **our defense** like that
 b. That was some **eagles** football right there boys.

There are two avenues of generalization to explore — generalization over the **entity**, and generalization over **the event**.

Preliminary analyses reveal 4 levels of **generalization** of reference:

- **Player name** **TK87** big money !!!
- **Groups within team** Did not expect **their o-line** to bully **our defense** like that.
- **team names** That was some **eagles** football right there boys.
- **team+supporters** Maybe **we** will get a lucky penalty call too .

Are there systematic differences in how commenters **refer** to in-group and out-group?

Short-term How does intergroup identity and state-of-the-world influence generalization over events and entities?

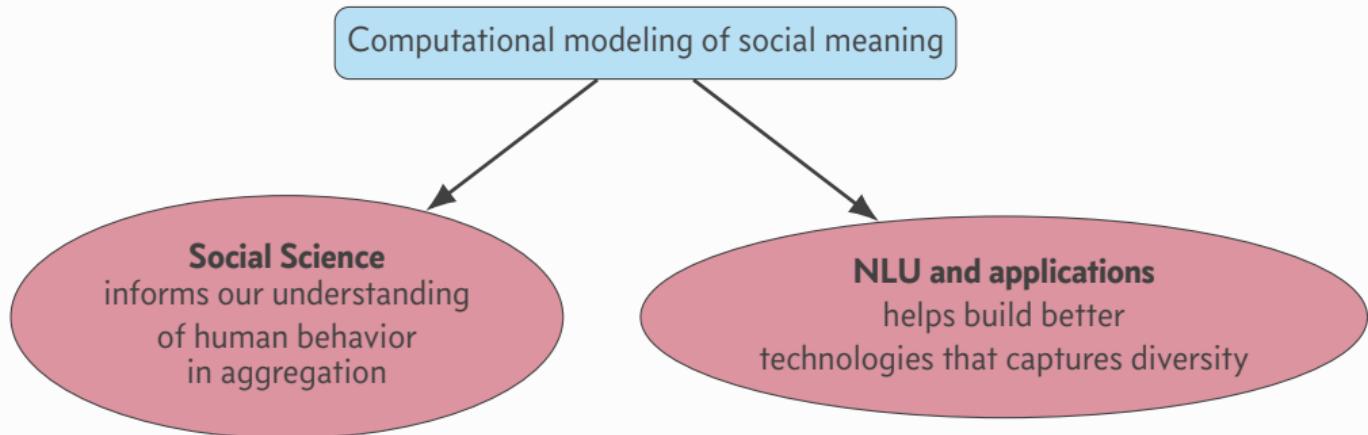
Medium-term What is the (intergroup) perspective in synthesized text from LLMs? How do they generalize over events and entities, and how is it different from humans?

Long-term Modern schema informed by theory and corpora to model generalizations with a focus on the influence of social identity.

Generalization is key to building systems with robust commonsense reasoning and knowledge.

We need to first understand the myriad ways we do it.

WHY MODEL SOCIAL MEANING?



COLLABORATORS

Jessy Li

David Beaver

Kyle Mahowald

Katrin Erk

Malihe Alikhani

Kate Atwell

Yuki Zang

Barea Sinno

Benjamin Chen

Rebecca Warholic

FIN

Thank you

REFERENCES I

- Beltrama, Andrea (2020). "Social meaning in semantics and pragmatics". In: *Language and Linguistics Compass* 14.9, e12398.
- Eckert, Penelope (2012). "Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation". In: *Annual review of Anthropology* 41, pp. 87–100.
- Hall-Lew, Lauren, Emma Moore, and Robert J. Podesva (2021). "Social Meaning and Linguistic Variation: Theoretical Foundations". In: *Social Meaning and Linguistic Variation: Theorizing the Third Wave*. Ed. by Lauren Hall-Lew, Emma Moore, and Robert J. Editors Podesva. Cambridge University Press, pp. 1–24.
- Govindarajan, Venkata S, Katherine Atwell, Barea Sinno, Malihe Alikhani, David Beaver, and Junyi Jessy Li (May EACL 2023). "How people talk about each other: Modeling Generalized Intergroup Bias and Emotion". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2488–2498.

- Govindarajan, Venkata S, Benjamin Chen, Rebecca Warholic, Katrin Erk, and Junyi Jessy Li (Nov. EMNLP 2020). "Help! Need Advice on Identifying Advice". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 5295–5306.
- Govindarajan, Venkata S, David Beaver, Kyle Mahowald, and Junyi Jessy Li (July Findings@ACL 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.
- Austin, John Langshaw (1975). *How to do things with words*. Vol. 88. Oxford university press.
- Zellers, Rowan, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi (June 2021). "TuringAdvice: A Generative and Dynamic Evaluation of Language Use". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4856–4880.

- Abolfathiasl, Hossein and Ain Nadzimah Abdullah (2013). "Pragmatic Strategies and Linguistic Structures in Making 'Suggestions': Towards Comprehensive Taxonomies". In: *International Journal of Applied Linguistics and English Literature* 2.6, pp. 236–241. issn: 2200-3452.
- Negi, Sapna, Tobias Daudert, and Paul Buitelaar (June 2019). "SemEval-2019 Task 9: Suggestion Mining from Online Reviews and Forums". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 877–887.
- Potamias, Rolandos Alexandros, Alexandros Neofytou, and Georgios Siolas (June 2019). "NTUA-ISLab at SemEval-2019 Task 9: Mining Suggestions in the wild". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 1224–1230.

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
- Alammar, Jay (2020). *The Illustrated BERT, ELMo, and co. (Blog post)*.
- 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).
- Plutchik, Robert (2001). "The Nature of Emotions". In: *American Scientist* 89.4, pp. 344–350.
- Wang, Sida and Christopher Manning (July 2012). "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 90–94.
- Mohammad, Saif M. and Peter D. Turney (2013). "Crowdsourcing a Word-Emotion Association Lexicon". In: *Computational Intelligence* 29.

- Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen (Oct. 2020). "BERTweet: A pre-trained language model for English Tweets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 9–14.
- Zad, Samira, Joshuan Jimenez, and Mark Finlayson (Aug. 2021). "Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon". In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online: Association for Computational Linguistics, pp. 102–113.
- Louis, Annie and Ani Nenkova (June 2011). "Text Specificity and Impact on Quality of News Summaries". In: *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, pp. 34–42.
- Li, Junyi Jessy (2017). "From Discourse Structure To Text Specificity: Studies Of Coherence Preferences". PhD thesis. University of Pennsylvania.

- Gao, Yifan, Yang Zhong, Daniel Preoțiuc-Pietro, and Junyi Jessy Li (July 17, 2019). "Predicting and Analyzing Language Specificity in Social Media Posts". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1. Number: 01, pp. 6415–6422. issn: 2374-3468.
- Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng (Nov. 2019). "The Woman Worked as a Babysitter: On Biases in Language Generation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3407–3412.
- Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng (Nov. 2020). "Towards Controllable Biases in Language Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 3239–3254.

ADVICE DATASET

We chose sentences as the unit of advice, and collected annotations for which sentences in a reply **contained advice**.

r/AskParents 10,594 sentences 407 posts

r/needadvice 7,862 sentences 277 posts

ADVICE ANNOTATION

1 Reply 1

3 In Kentucky it 's legal to leave a " mature " 8 year old at home alone all day .

4 I find that crazy young .

8 I started leaving mine home at age 9 - 10 for a half hour here , 45 min there , working up to a couple of hours .

10 >>> Reply 1.1

12 Yeah , even though my son has always been very mature for his age , I would not have been comfortable leaving him home alone all day long at age eight !

14 -----

16 Reply 2

18 For an hour ?

19 I 'd on average say elementary school aged .

20 So 6 up , depending on how responsible / mature the child is and if they 're willing to stay home alone .

22 No answering the door , no leaving the house , no using the stove , no friends over and I 'd talk about what neighbors might be home in case of an emergency .

24 Oh , and I 'm from Germany .

ADVICE ANNOTATION

1 Reply 1

Advice

3 In Kentucky it 's legal to leave a " mature " 8 year old at home alone all day .

4 I find that crazy young .

Advice

8 I started leaving mine home at age 9 - 10 for a half hour here , 45 min there , working up to a couple of hours .

10 >>> Reply 1.1

Advice

12 Yeah , even though my son has always been very mature for his age , I would not have been comfortable leaving him home alone all day long at age eight !

14

16 Reply 2

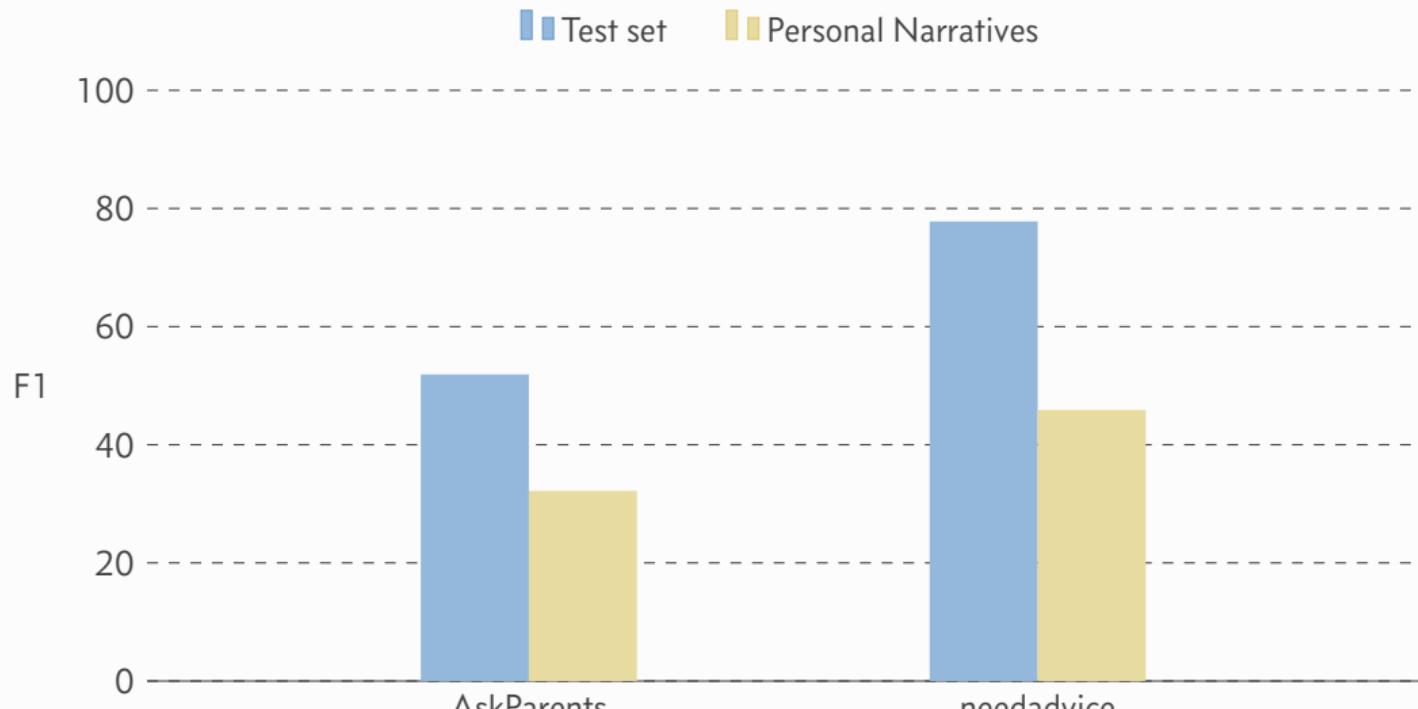
Advice

18 For an hour ? I 'd on average say elementary school aged . So 6 up , depending on how responsible / mature the child is and if they 're willing to stay home alone .

20 No answering the door , no leaving the house , no using the stove , no friends over and I 'd talk about what neighbors might be home in case of an emergency .

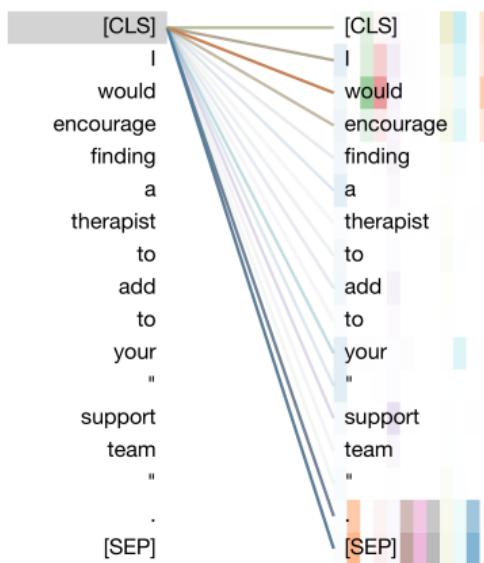
22 Oh , and I 'm from Germany .

PERFORMANCE ON PERSONAL NARRATIVES



Best model's performance on personal narrative sentences in test set suffers.

NEURAL MODELS CAN BE INSIGHTFUL



'Attention' patterns of a neural model trained to predict if a sentence contains advice. Colors reveal higher probability mass assigned to input tokens by parameters in the neural model.

ANNOTATION PROCEDURE

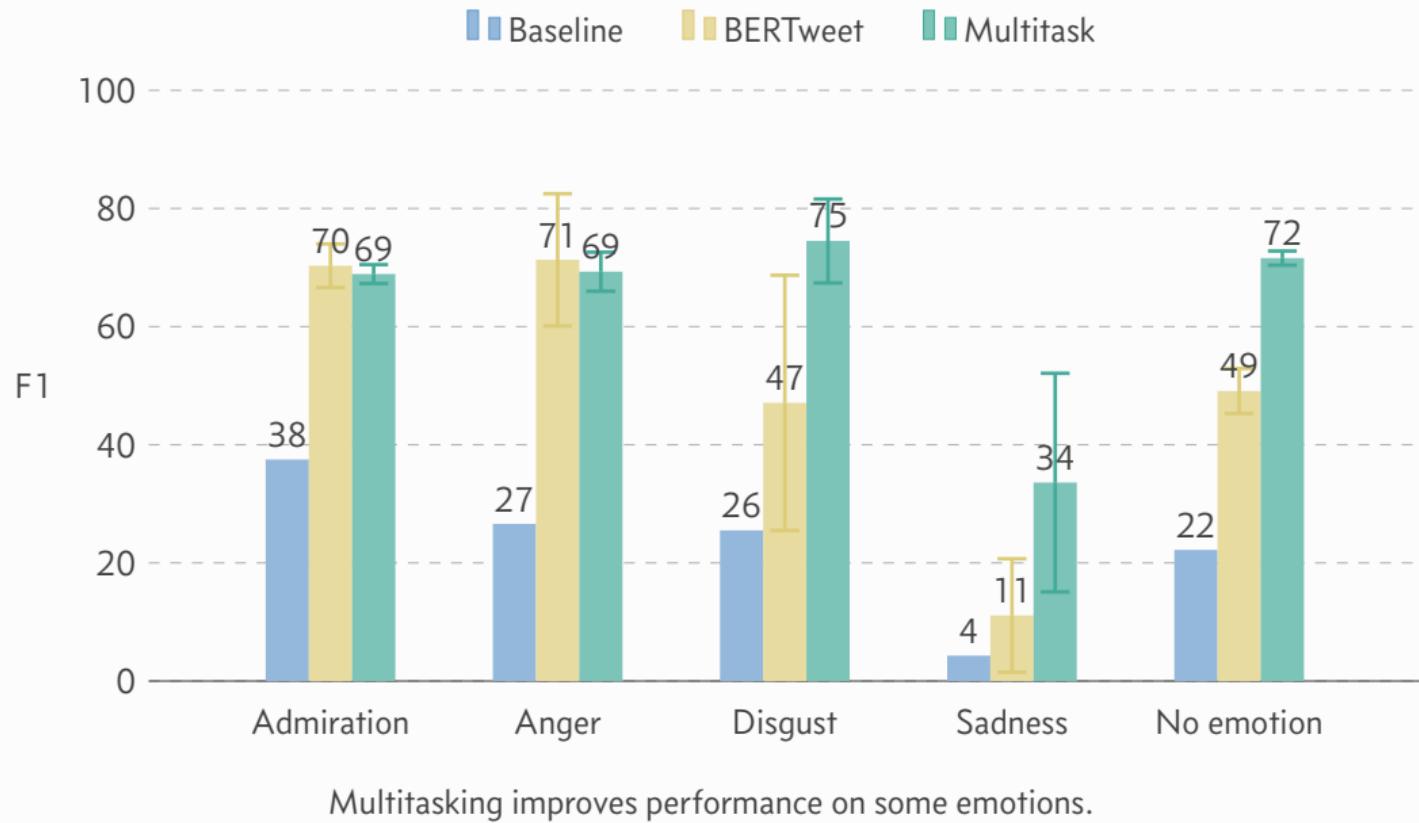
Example Tweet

If @Doe can get her hair done in person, Congress can vote in person. Further, if @JoeBiden can vote in person, Americans should be encouraged to cast their vote in person.

Please select only the **most notable emotions** you think are expressed by writer in connection with @Doe in the tweet.

Fear, Admiration, Joy, Interest, Anger, Disgust, Sadness, Surprise

RESULTS-EMOTION



ERROR ANALYSIS

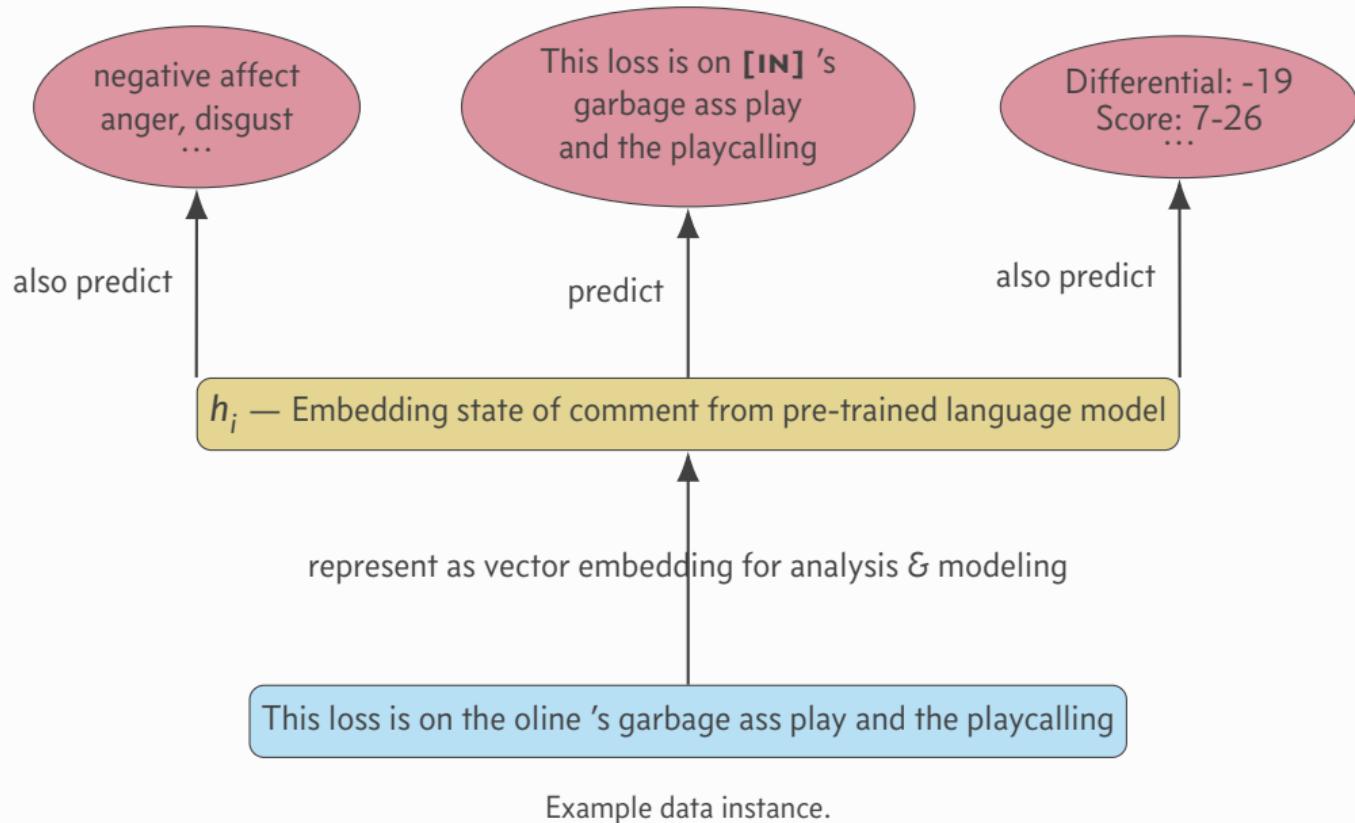
Models **beats trained annotators** in some cases — annotators fall back on the heuristic that positive emotions *probably* mean in-group tweet, but bipartisanship displays are common in US Congress:

- ⑯ a. Admire @OfficialCBC Chairman **@Doe**'s moral voice on issues of racism and restorative justice...
- b. Proud to work with **@Doe** to #ReviveCivility. #tbt Read more about our efforts here...

The model still makes basic errors though:

- ⑯ Trump selected **@Doe** for HHS Secretary. Price has undeniable history of cutting access to healthcare to millions, especially women.

MULTITASKING



... a socially acceptable outlet for xenophobia. That is the function of organized sports (in society) for the most part...

John Siracusa on Reconcilable Differences Ep.3

Jessy's Prompt

Suppose you are a supporter of the **French football team** at the 2018 World Cup. How would you describe the final?

ChatGPT output

If I were a supporter of the French football team at the 2018 World Cup, I would describe the final as a thrilling and historic victory for France. The final match was played on July 15, 2018, at the Luzhniki Stadium in Moscow, where France faced off against Croatia. France won the game with a score of 4-2, earning their second World Cup title and their first since 1998.

The final was a high-scoring and exciting game, with both teams displaying impressive skills and tactics. France took the lead early in the game with an own goal by Croatia, and then **added to their lead with goals by Antoine Griezmann, Paul Pogba, and Kylian Mbappe**. Croatia fought back with goals by Ivan Perisic and Mario Mandzukic, but ultimately fell short against a strong French team.

The victory was especially meaningful for France, as it was a moment of national pride and unity. The team's **diverse and multicultural makeup** reflected the country's values of inclusion and integration, and their success on the world stage was a source of joy and celebration for fans all over France and beyond. Overall, the final was an unforgettable moment in football history, and a testament to the skill, teamwork, and spirit of the French team.

Jessy's Prompt

Suppose you are a supporter of the **Croatian football team** at the 2018 World Cup. How would you describe the final?

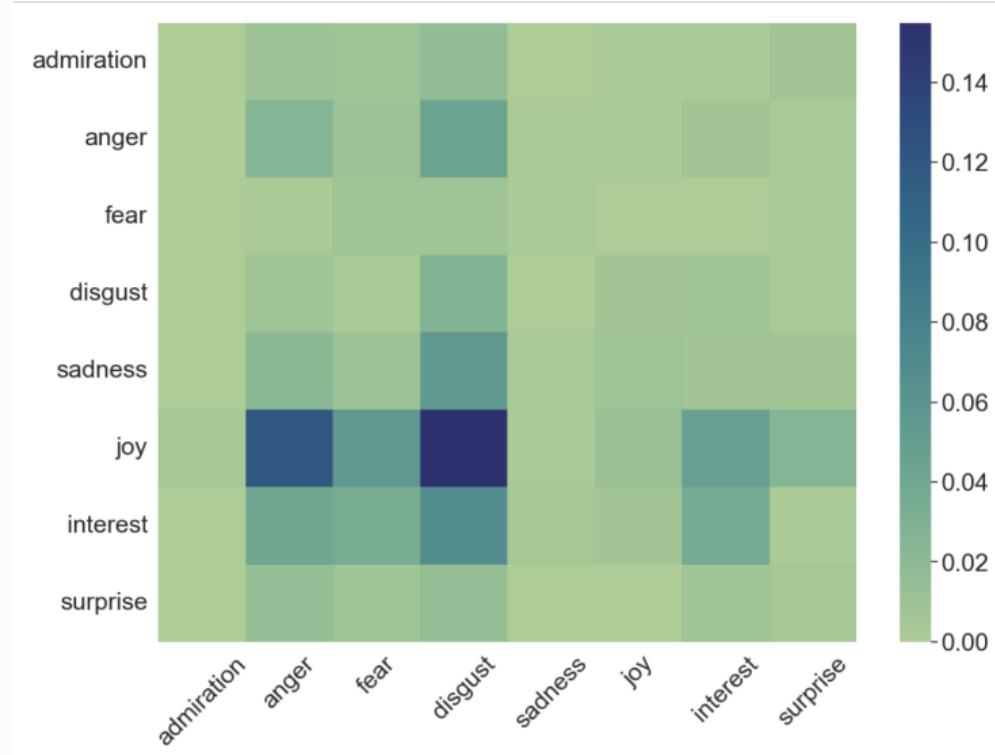
ChatGPT output

If I were a supporter of the Croatia football team at the 2018 World Cup, I would describe the final as a disappointing loss, but still a source of immense pride and accomplishment for our team. The final match was played on July 15, 2018, at the Luzhniki Stadium in Moscow, where Croatia faced off against France. France won the game with a score of 4-2, but **the scoreline does not reflect the level of competition and intensity** of the game.

Despite the loss, the Croatia team played with passion and determination throughout the tournament, defeating strong teams like Argentina, England, and Russia on their way to the final. The final itself was a hard-fought battle, with Croatia falling behind early due to an own goal, but fighting back with a goal by Ivan Perisic to equalize. Unfortunately, France proved to be too strong on the day, and scored two more goals to seal the victory.

However, despite the outcome of the game, the Croatia team's performance throughout the tournament was a source of pride and inspiration for our country. **Our team showed great skill, resilience, and teamwork, and their journey to the final was a testament to the talent and dedication of Croatian football.** Although we did not ultimately win the tournament, our team's accomplishments in 2018 will always be remembered and celebrated by Croatian fans around the world.

ALIGNED EMOTIONS



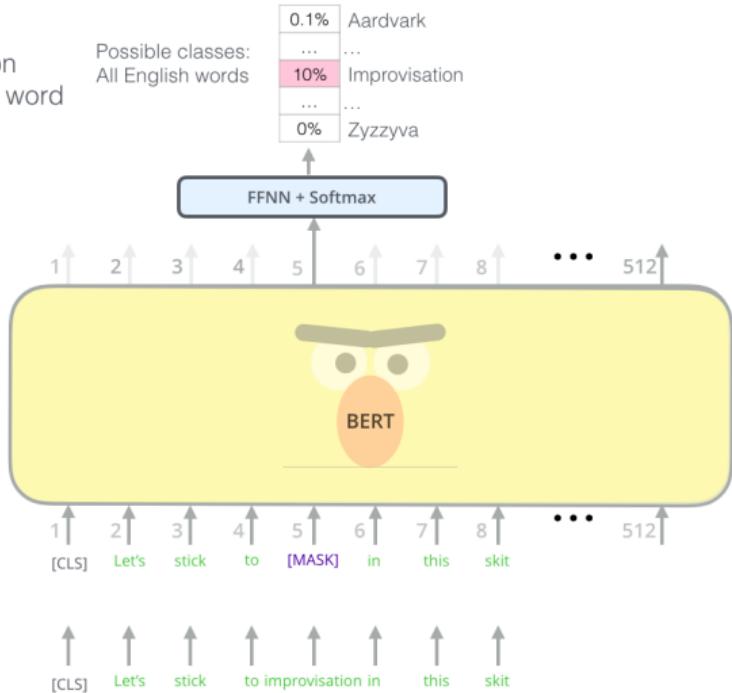
Aligned emotions **towards** the Chiefs from Chiefs (y-axis) and Eagles(x-axis)

WHAT IS BERT?

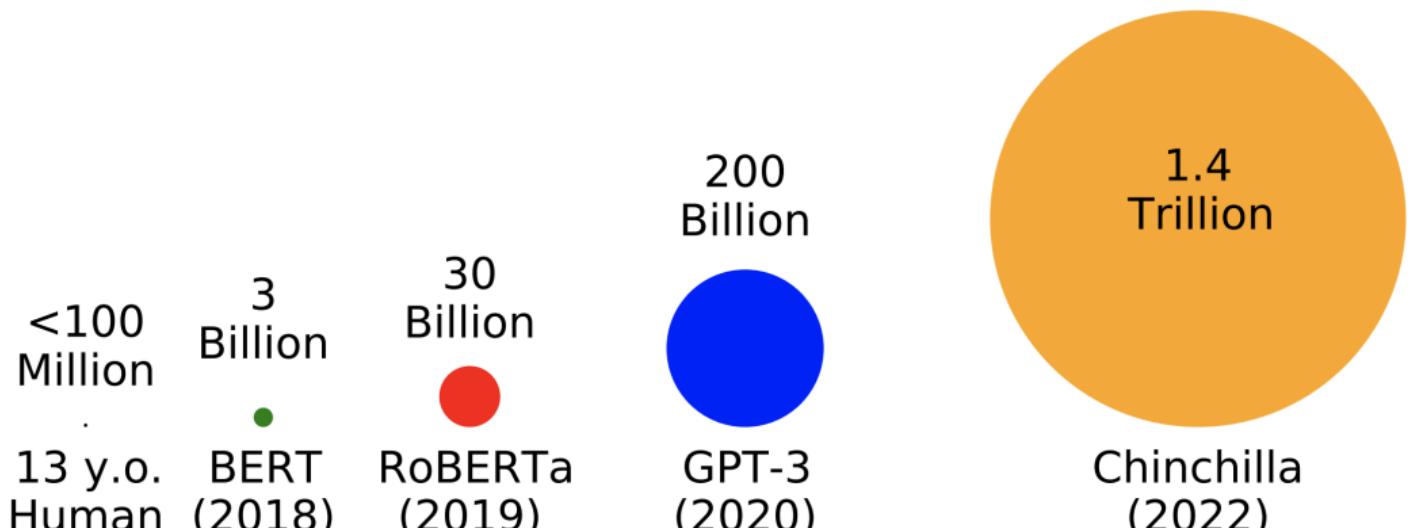
Use the output of the masked word's position to predict the masked word

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zzyzyva

Randomly mask 15% of tokens



Through pre-training on word prediction on large corpora, neural language models compress and **encode** a large amount of linguistic information. Illustration from *The Illustrated BERT, ELMo, and co. (Blog post)* (2020)



Language models are not sample efficient.

LIL BEVO



We explore 3 strategies **inspired by cognition** for a sample efficient Language Model:

- Is there structure in music that helps model learn structure in language?
- Curriculum training with shorter sequences first.
- Targeted linguistic tuning.

TAKEAWAY

Unsurprisingly, these models **don't learn like humans**. We should understand their 'learning' in a conceptually different manner.