Computational Linguistic Models of Social Identity

AND OTHER PROJECTS

Venkat

Feb 6 2024

The University of Texas at Austin

How does our **social identity** influence the language we use?

How can we use **machine learning** to understand human communicative behavior?

How do we build better models that understand language infused with meaning from social identity?

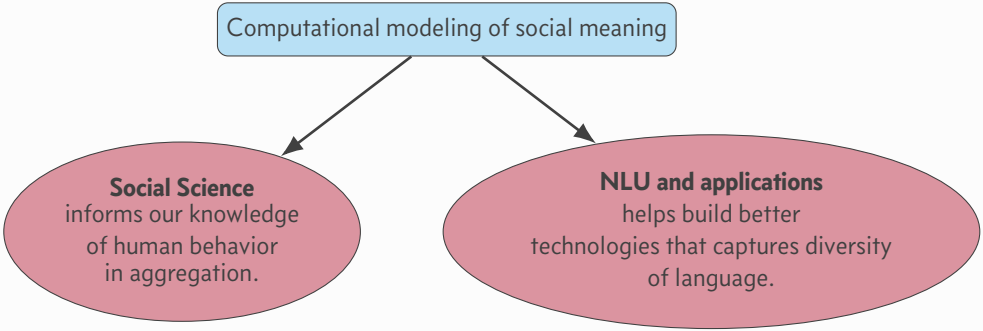
…the constellation of qualities and properties that linguistic forms convey about the social identity of language users

Beltrama, 2020

All communicative behavior is **situated in social context** and is informed and moulded by it.

Eckert, 2012; Hall-Lew et al., 2021

WHY MODEL SOCIAL MEANING?



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.

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

Top word 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)

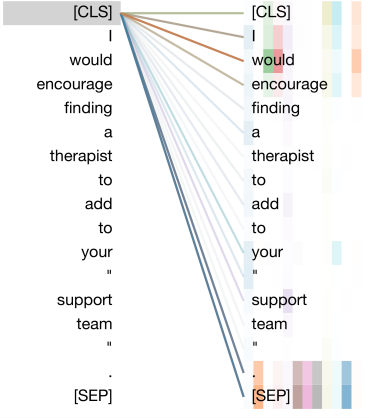
② a. I **would encourage** finding a therapist to add to your support team.

b. **I talked** on Reddit with others to get support and ideas .

Both utterances give advice — the latter is more implicit.

Computational models can uncover diverse ways of doing things with language, and be informative even when they fail.

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.

As I will show in "Help! Need Advice on Identifying Advice"(EMNLP 2020), we can make inferences from errors in trained neural models as well.

OUTLINE OF TALK

① How does our social identity help us **do things** with language? EMNLP 2020 paper

② How does **in-group versus out-group** identity unconsciously modify our speech? EACL 2023 paper

③ Other research projects building **efficient** machine learning models, understanding how models work, and my personal interest in building **focused applications** using modern machine learning.

UNDERSTANDING HOW PEOPLE GIVE ADVICE ONLINE

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

③ 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?

Austin, 1975; Zellers et al., 2021

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

How is advice **structured** online?

We introduce a dataset of online advice, showing various strategies and the subtle influence of social identity towards giving implicit advice.

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

Specifically, how good are models at identifying implicit advice conveyed through subtle cues?

Parenting with a history of depression?

④ **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 datasets 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

16.4% of advice sentences from r/AskParents (and **7%** from r/needadvice) were judged to contain **personal narratives** from a sample analysis.

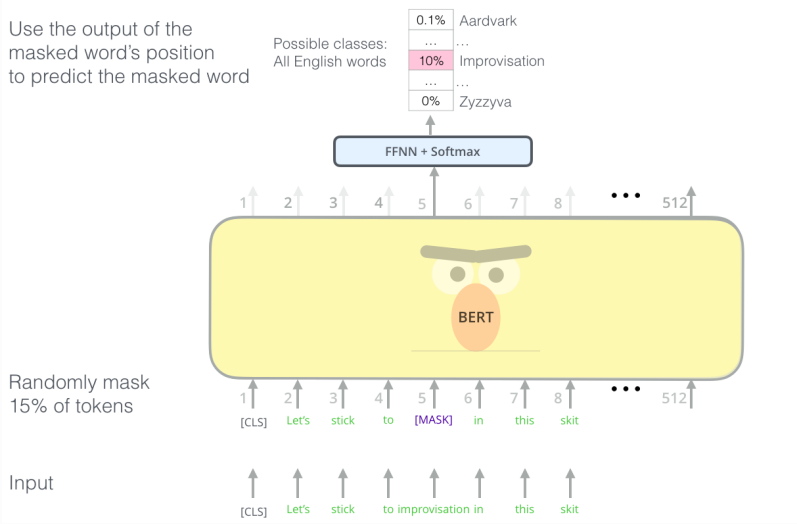
How well do models identify advice, especially advice from personal narratives?

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

Rule-based Match and score against **words**, **phrases**, **regexs**:
suggest, *recommend*, *.*would\slike.*if.**

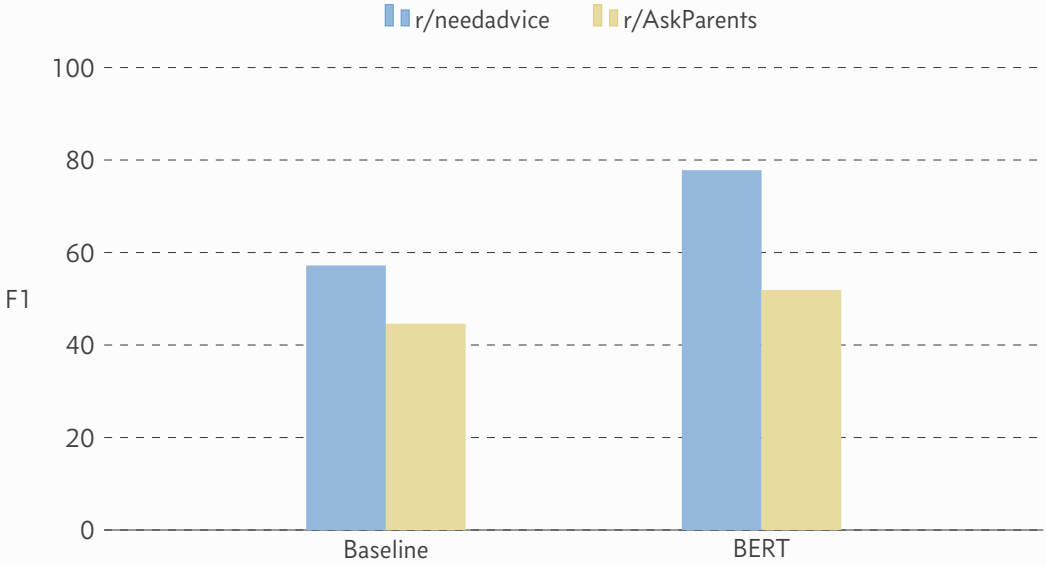
Learned embeddings We use **BERT** and perform binary classification with a sentence-level representation.

Negi et al., 2019; Potamias et al., 2019; Devlin et al., 2019



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.

What about unintentional things that we do with language?

Like **implicit bias** when we talk about someone else.

IN-GROUP VS. 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**.

Maass, 1999; Gorham, 2006

White participants saw a news reel with a white/black criminal suspect and were then asked to describe what they saw in the news reel:

⑤ 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**.

White participants were more likely to describe the event with abstract words like in (c) and (d) **when the suspect was black (out-group)**.

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

⑥ 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 Doe — (a) is **in-group**, (b) is **out-group**. Notice the word *bipartisan* in (b), a subtle indicator of bias in this dimension.
- the intensity of admiration expressed by the speaker towards Doe is greater in (a).

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?

- Tweets by members of US Congress which mention one other member.

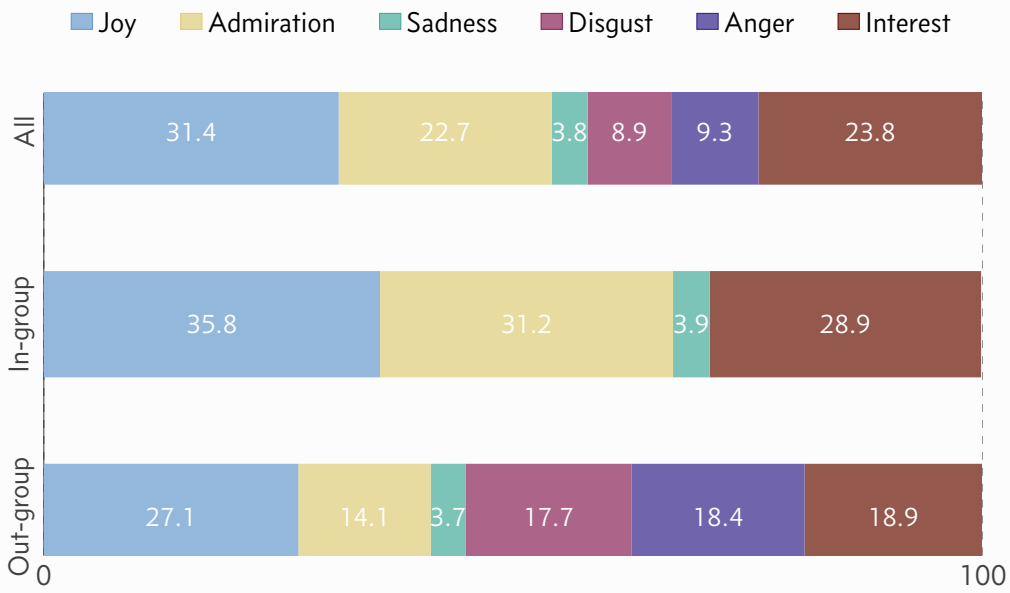
- Tweets are either directed in-group or out-group.

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

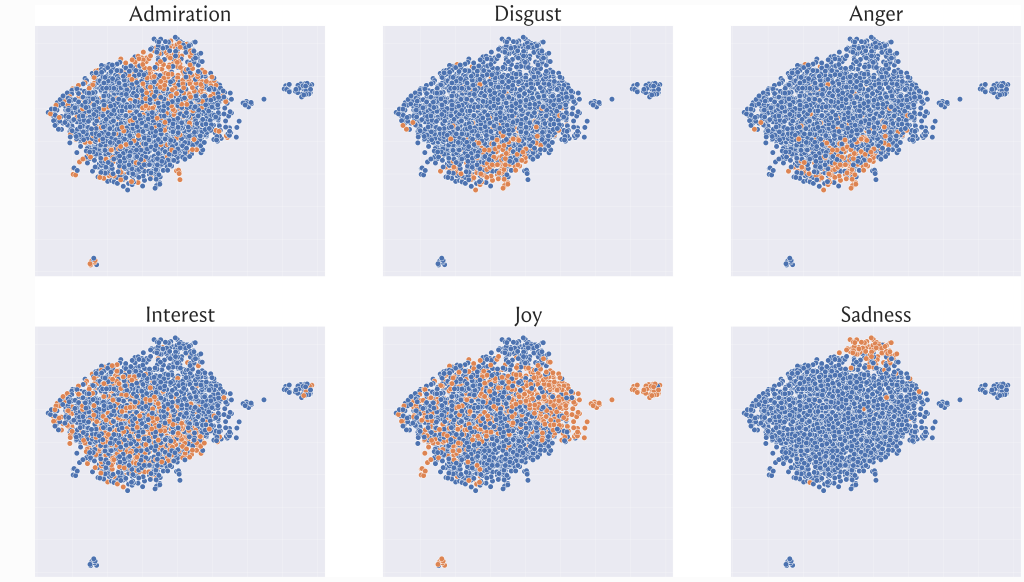
GITHUB.COM/VENKATASG/INTERPERSONAL-BIAS

Plutchik, 2001

EMOTION DISTRIBUTION



TWEET EMBEDDINGS & GOLD EMOTIONS



Tweet embeddings from a language model projected using UMAP. Each point is a tweet and orange indicates the emotion is present. Observe the separability of clusters of emotions.

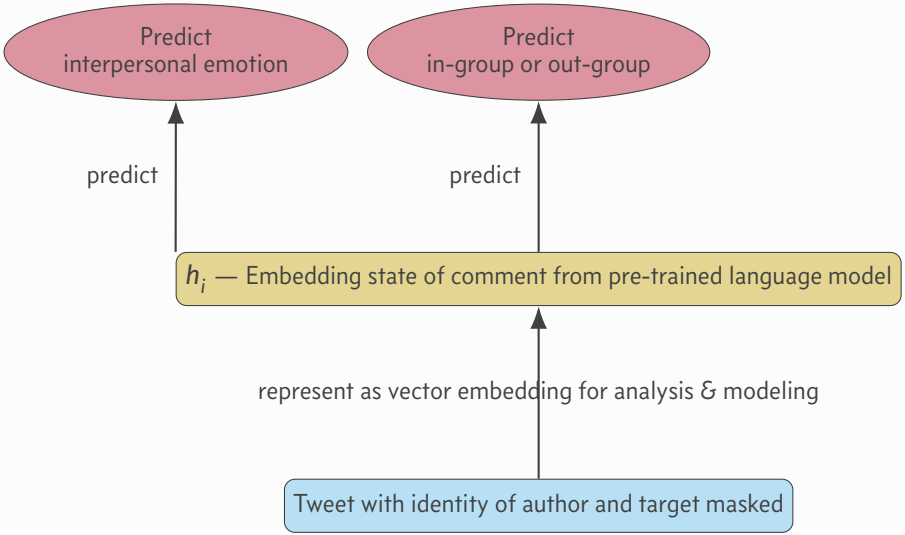
Two tasks: predict Intergroup Relationship and Interpersonal Emotion.

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

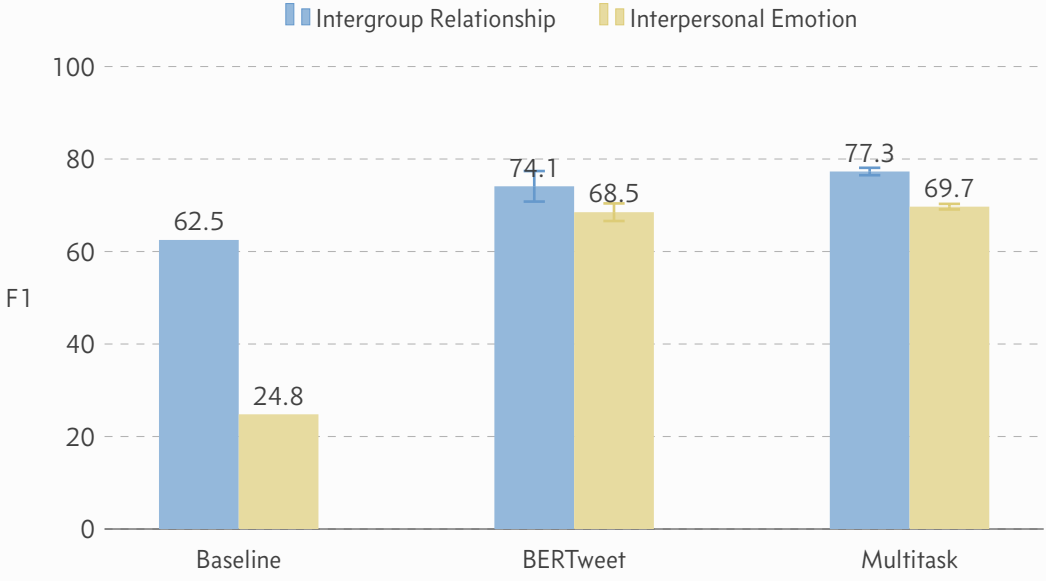
BERTweet Predict both dimensions with classification or labelling layer on top of finetuned BERTweet embeddings.

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

Wang et al., 2012; Mohammad et al., 2013; Nguyen et al., 2020; Zad et al., 2021



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



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.

⑦ 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**.

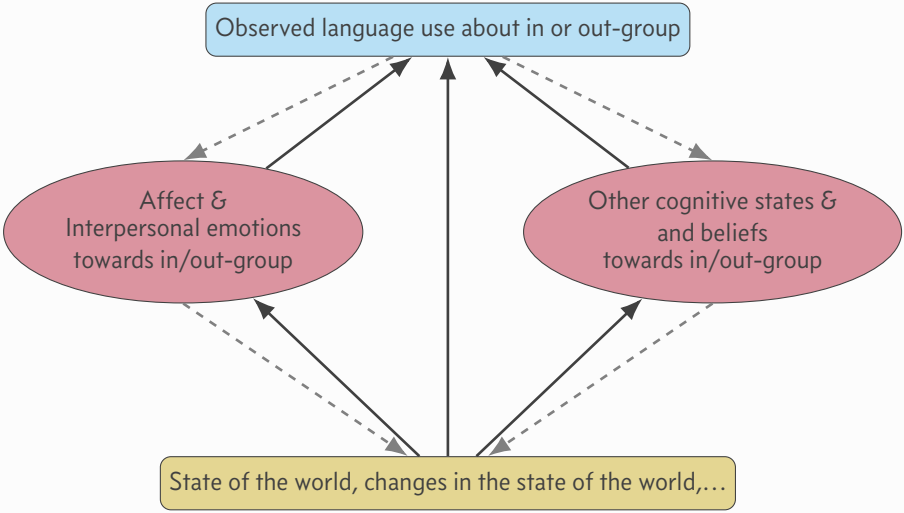
	Socially desirable	**Socially undesirable**
in-group	abstract	concrete
out-group	concrete	abstract

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.

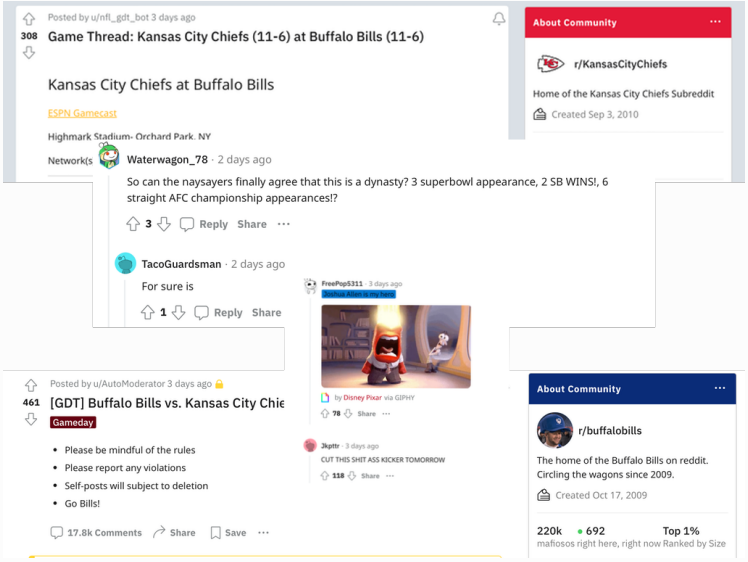
Can we do better?

ONGOING WORK, AND MOVING FORWARD



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

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.



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?

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

Understand how references to in-group and out-group vary with game state (score), and how people generalize over team, and play.

Next steps Diversity across dialects and languages with aligned events.

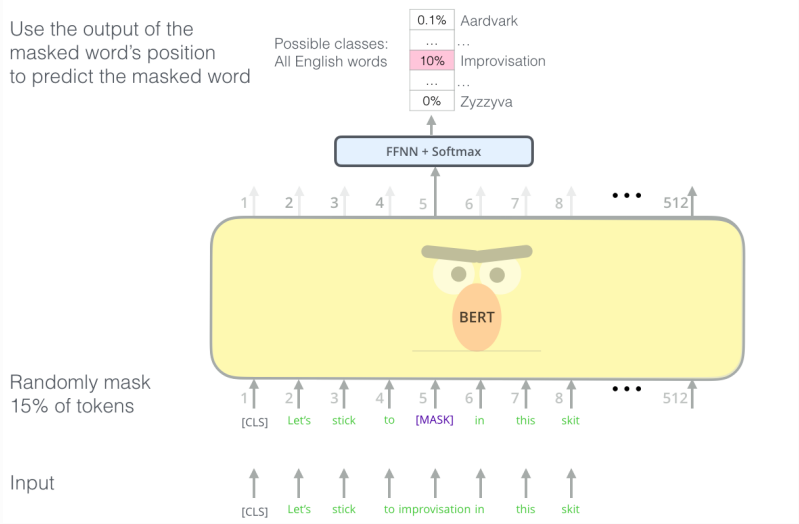
English commentator	Play	Spanish commentator
Gano from 63 yards out..it is good!	Field goal by Gano	GANO LO GANO! GANO LO GANO! GANO LO GANO!!

Long-term

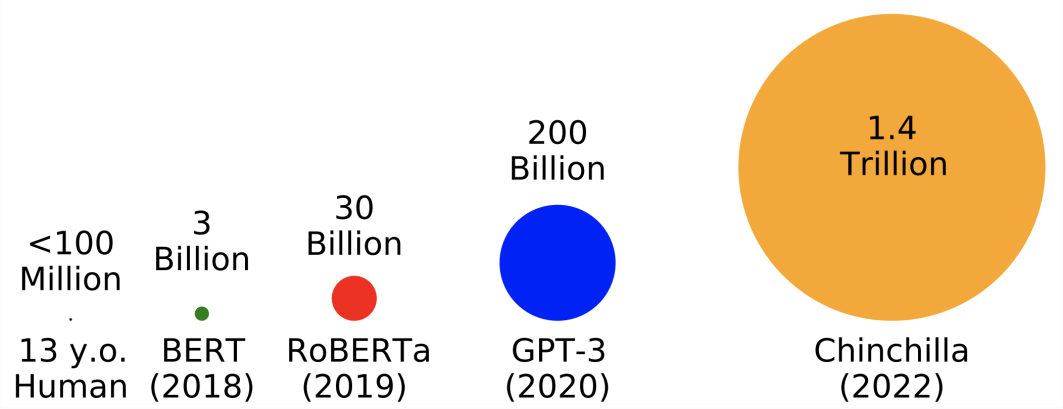
- Social meaning variation in non-polarized settings.
- Modeling intersectional social identity.
- Intergroup perspective in Large Language Model (LLM) outputs.

OTHER RESEARCH PROJECTS AND INTERESTS

Train and build **efficient** language models.



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)

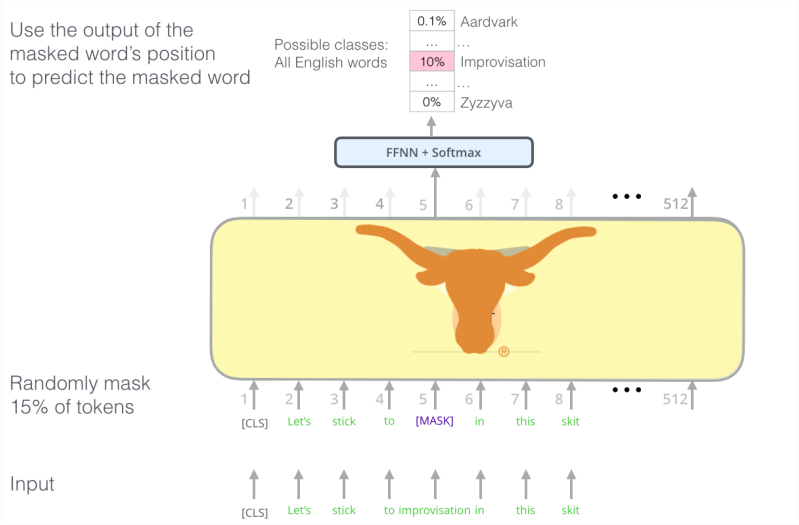


Number of words different model are exposed to during training, compared to average number of words heard by a 13 year old.

We build our small language model **Lil Bevo** — named after UT Austin's mascot.

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

- Train on **piano music** first before training on language data — is there structure in music that helps learn structure in language?
- Curriculum training with shorter sequences first — easing the model into learning.
- Targeted linguistic tuning — finding examples of specific linguistic phenomenon ()

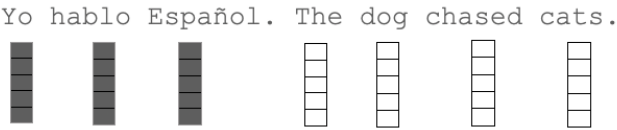


Bevo is an encoder language model just like BERT, we train it differently.

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

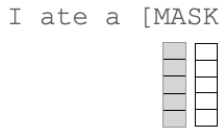
Understanding **internals** of how machine learning models work.

(1) Lang id classifier trained on English vs. Spanish



(2) AlterRep applied to a masked token from English sentence, to push it towards Spanish.

(3) Compare log prob of control to target/ random answers from English, Spanish.



English Target	English Random	Spanish Target	Spanish Random	Control (Hindi random)
cherry	book	cereza	lapiz	संगणक

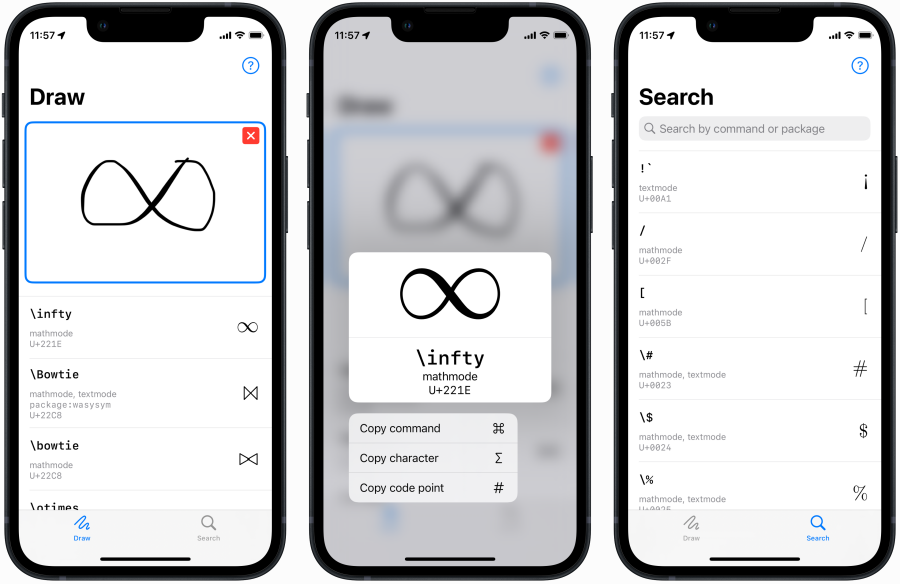
Do model internal representations have structure we can exploit, say for doing **translation**?

Most likely tokens pre-intervention	friend, house, dream, novel, room, bed, book
Most likely tokens after pushing to Spanish	coma, car, man, la, son, del, más
Most likely tokens after pushing to English	house, dream, room, friend, book, tree, memory

Example of the most likely tokens predicted for the masked token pre and post-intervention for the English language text “One day while Cat was wandering about, he came to a ….”

Controlled interventions push model predictions to the prior over target language.

Build focused apps that use machine learning in **efficient**, **targeted** ways to do one thing well.



Screenshots of my app DeTeXt running on iPhone.

- Train a neural network on dataset of LaTeX symbols and drawing co-ordinates.
- Reduce the size of the model and . Trade-off a little accuracy for lots of gains in efficiency. Final neural network is only **15 megabytes**.
- Incorporate into a well-designed user interface (this is the hard part).
- Source code for app is available online: github.com/venkatasg/detext.

Social meaning Social identity is crucial to meaning and language use, and computational modeling can enrich our understanding of its influence and importance.

Machine Learning How do we build efficient machine learning models for language understanding? How do we **interpret** how these models work?.

Apps Building focused apps that use machine learning efficiently to do one thing (or a few things) well.

FIN

Thank you

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.

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)*.

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.