How People talk about each other: Modeling Intergroup Bias and Emotions

Venkat
venkatasg@utexas.edu
The University of Texas at Austin

COLLABORATORS

- Katherine Atwell[1]
- Malihe Alikhani[1]
- Barea Sinno[2]
- David I. Beaver[3]
- Junyi Jessy Li[3]

[1] University of Pittsburgh
[2] Rutgers University
[3] The University of Texas at Austin

How do we study social bias in communication?

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. **Language is biased one way or another**.

How do we bring this insight into our work?

Van Dijk 2009, Beaver et al. 2018, Kaneko et al. 2019, Sheng et al. 2019, Sap et al. 2020, Webson et al. 2020, Pryzant et al. 2020, Sheng et al. 2020

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.

Maass 1999, Gorham 2006

- Wide variety of interpersonal utterances beyond elicited utterances in experiments.
- Restrictive conditions under which the LIB has been proven to exist — high polarization, with topic confined to those on which stereotypes exist.
- LIB focuses only on the abstractness of the predicate, and most studies are hand-coded.

Our approach

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

② We stand w @Doe, who has seen a lot worse than cheap insults from an insecure bully. #MLK-DAY weekend.

③ Parents and families live in constant fear for their children with food allergies. A worthy bipartisan cause - thank you @Doe for your leadership on this issue.

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

- the relationship between speaker and Doe — ② is **in-group**, ③ is **out-group**. Notice the word *bipartisan* in ③, a subtle indicator of bias in this dimension.
- the intensity of admiration expressed by the speaker towards Doe is greater in ②.

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

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

DATA & PRELIMINARY ANALYSIS

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.

DATA REQUIREMENTS

- Utterances which are directed at or are about another individual.
- Relationship between speaker and target known.
- Can be easily annotated for interpersonal emotion.

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

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

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

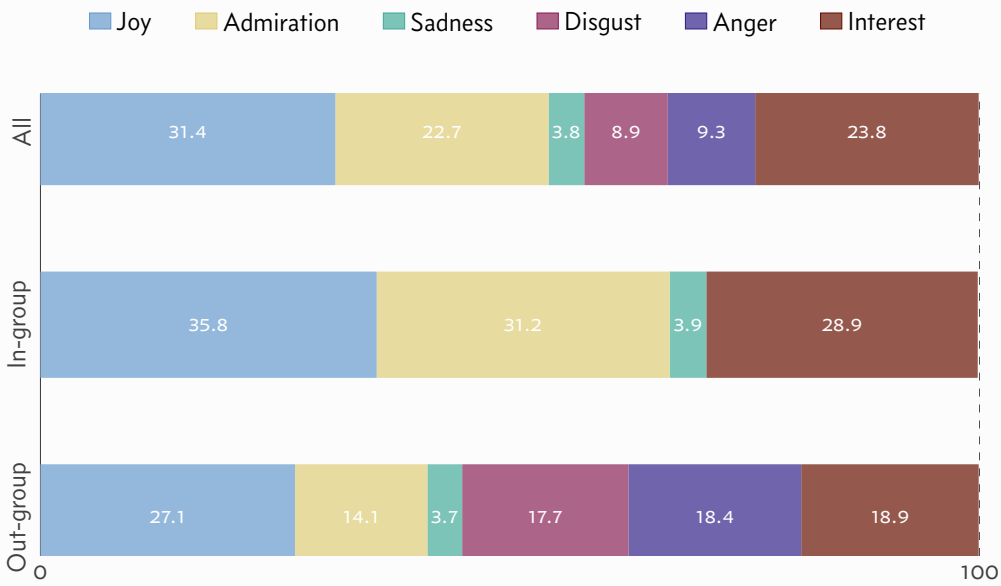
Plutchik 2001

Example Tweet

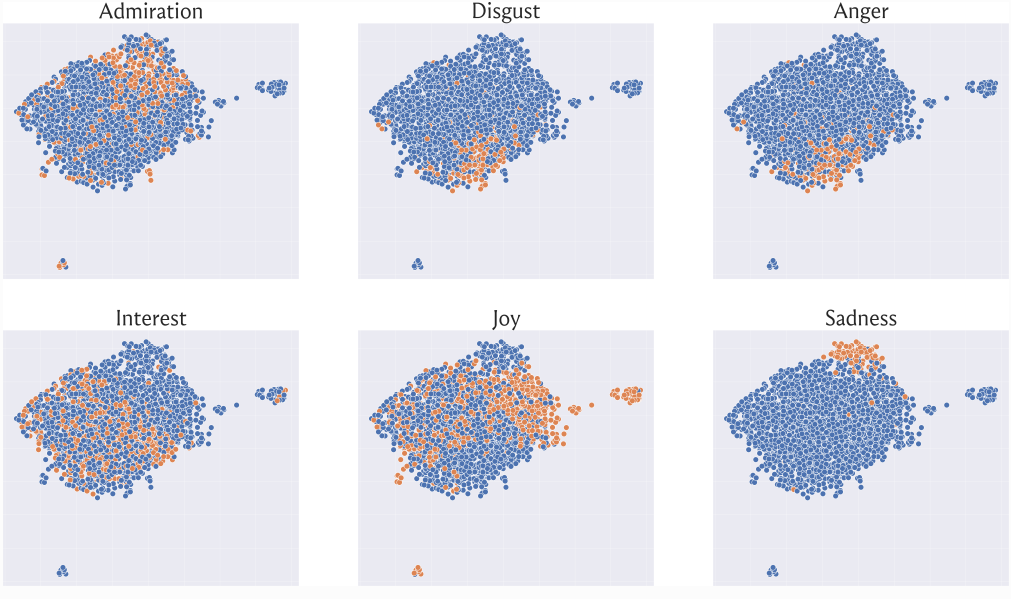
If **@Doe** can get her hair done in person, Congress can vote in person. Further, if @JoeBiden can vote in person, Americans should be encouraged to cast their vote in person.

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

EMOTION DISTRIBUTION



TWEET EMBEDDINGS & GOLD EMOTIONS



Tweet embeddings projected using UMAP. Each point is a tweet and orange indicates the emotion is present.

EXPERIMENTS

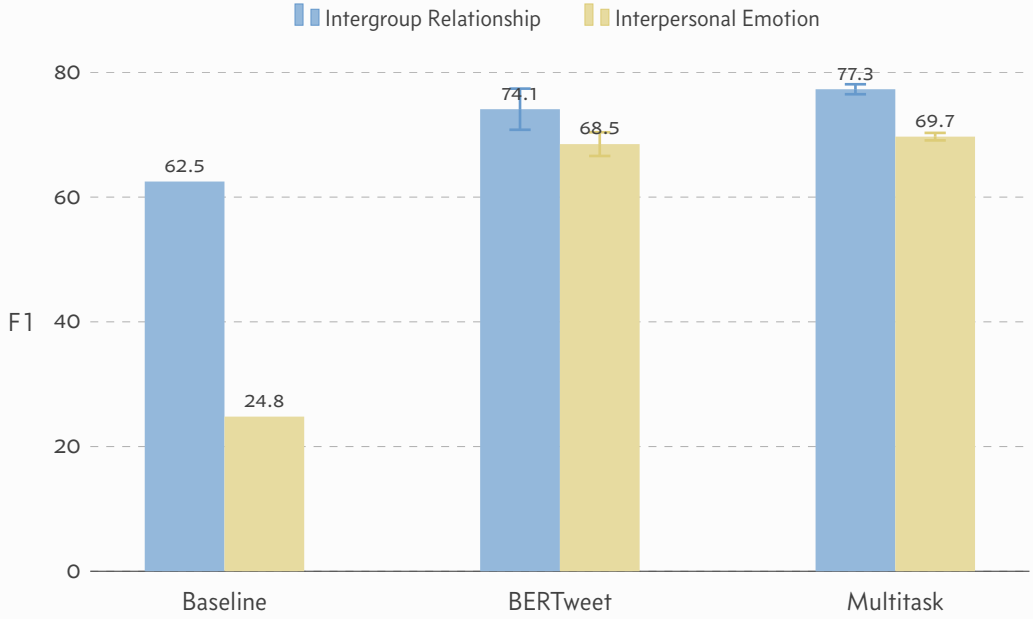
Two tasks: predict Intergroup Relationship and Interpersonal Emotion.

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

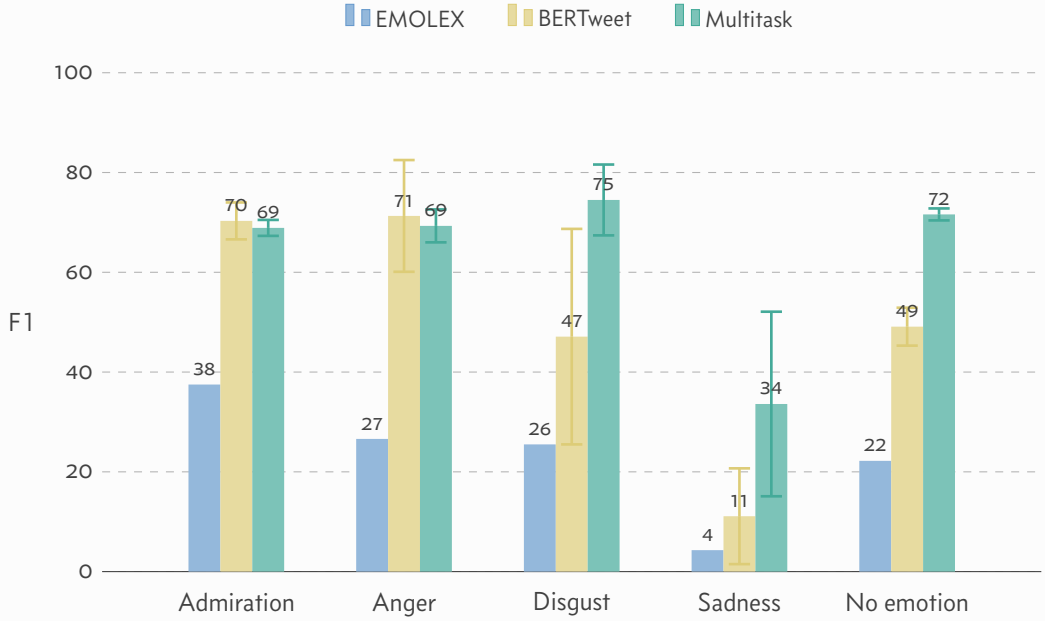
BERTweet Predict both dimensions with classification or labelling layer on top of finetuned BERTweet embeddings.

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


Intergroup Relationship
Interpersonal Emotion
F1
80
60
40
20
0
62.5
24.8
74.1
68.5
77.3
69.7
Baseline
BERTweet
Multitask

RESULTS-EMOTION



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:

④ 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:

⑤ Trump selected @Doe for HHS Secretary. Price has undeniable history of cutting access to health-care to millions, especially women.

SUMMARY

SUMMARY

Intergroup Bias Novel framing of bias based on interpersonal relationships — we situate interpersonal bias in intergroup relationship and interpersonal emotion.

Emotions Interpersonal emotions systematically varies with intergroup relationship context.

Multitasking Multitasking improves performance on both dimensions — fortifying the systematic interaction between the two.

Future Work What linguistic features underlie systematic variation in in-group and out-group language? How generalizable are the results to other domains with more situated data?

Fin.

Thank you.

Data & code available at:

github.com/venkatasg/interpersonal-bias

Links to paper & slides at: venkatasg.net/talks

INTER-ANNOTATOR AGREEMENT

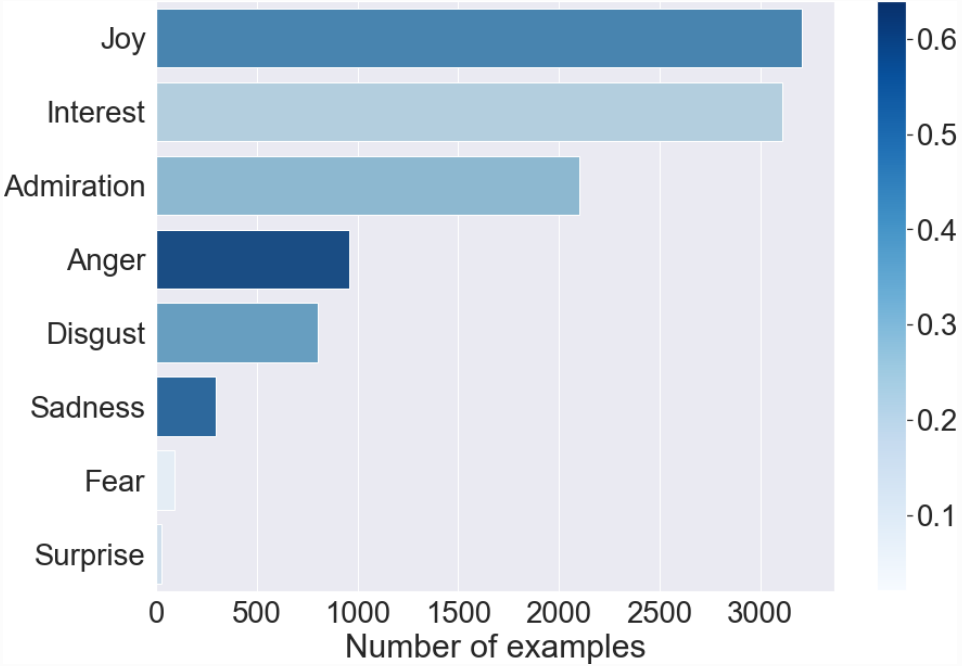
We measure **Plutchik Emotion Agreement** (PEA score) so that emotions that are closer together (like joy and admiration) are not penalized as highly as dissimilar emotions (like joy and sadness).

We find a PEA (min) of **0.6** and a PEA (max) of **0.73** indicating moderate to high agreement.

We also present inter-rater correlations for different emotions.

Desai et al. 2020, Demszky et al. 2020

INTER-RATER CORRELATIONS



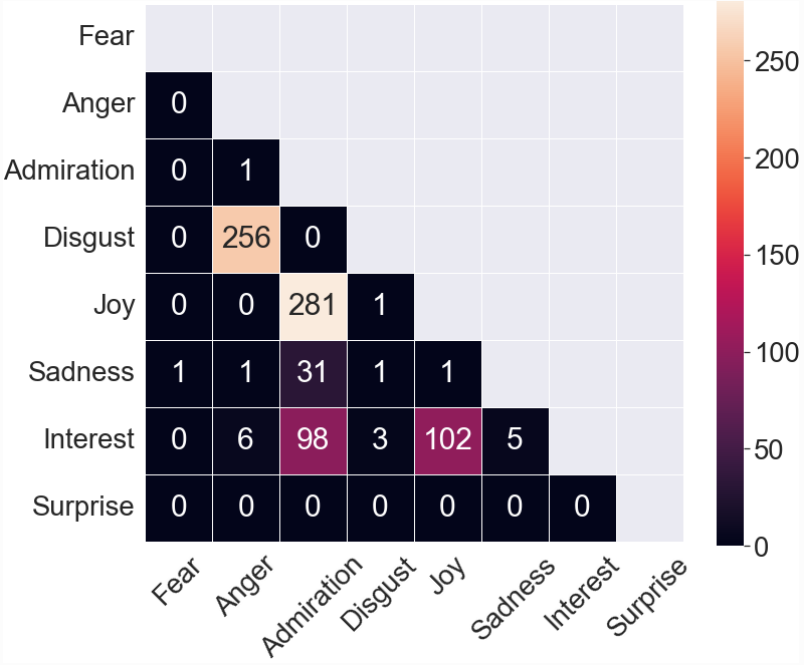
Emotion	Train	Dev	Test
Admiration	467	64	58
Anger	225	40	46
Disgust	206	32	43
Fear	1	0	0
Interest	701	83	84
Joy	801	107	106
Sadness	72	11	11
Surprise	1	0	0
No Emotion	519	56	63

Distribution of emotions in train-dev-test split

Emotion	All (%)	In-Group (%)	Out-Group (%)
Admiration	15.5	22.2	9.1
Anger	8.2	1.0	15.1
Disgust	7.4	0.3	14.2
Fear	0	0	0
Interest	22.9	27.2	18.6
Joy	26.7	32.2	21.4
Sadness	2.5	2.6	2.4
Surprise	0	0	0
No Emotion	16.8	14.5	19.1

Percentage of emotions in different interpersonal contexts

CO-OCCURENCE OF EMOTIONS



HUMAN LABELLING

- We investigate if human annotators were capable of accurately performing the IGR prediction task when the speaker and target are masked.

- Two authors of this paper, one a social science graduate student, and the other a computational linguistics graduate student, annotated 50 random tweets from our validation data which they had not been exposed to earlier for in/out group labels.

- Their Fleiss κ agreement score was **0.64**. Their scores on these 50 tweets were **0.67** and **0.63**, below peak model performance.

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

Top unigram and bigram features from NB-SVM model for each class.

Beaver, David & Jason Stanley. 2018. Toward a Non-Ideal Philosophy of Language. *Graduate Faculty Philosophy Journal* 39(2). 503–547.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade & Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. Online: Association for Computational Linguistics.

Desai, Shrey, Cornelia Caragea & Junyi Jessy Li. 2020. Detecting Perceived Emotions in Hurricane Disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5290–5305. Online: Association for Computational Linguistics.

Gorham, Bradley W. 2006. News Media's Relationship With Stereotyping: The Linguistic Intergroup Bias in Response to Crime News. *Journal of Communication* 56(2). Place: United Kingdom Publisher: Blackwell Publishing, 289–308.

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

Maass, Anne. 1999. Linguistic Intergroup Bias: Stereotype Perpetuation Through Language. In Mark P. Zanna (ed.), *Advances in Experimental Social Psychology*, vol. 31, 79–121. Academic Press.

Mohammad, Saif M. & Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29.

Nguyen, Dat Quoc, Thanh Vu & Anh Tuan Nguyen. 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*, 9–14. Online: Association for Computational Linguistics.

Plutchik, Robert. 2001. The Nature of Emotions. *American Scientist* 89(4). 344–350.

Pryzant, Reid, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky & Diyi Yang. 2020. Automatically Neutralizing Subjective Bias in Text. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(01). 480–489.

Sainburg, Tim, Leland McInnes & Timothy Q Gentner. 2021. Parametric UMAP Embeddings for Representation and Semisupervised Learning. *Neural Computation* 33(11). 2881–2907.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith & Yejin Choi. 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*, 5477–5490. Online: Association for Computational Linguistics.

Sheng, Emily, Kai-Wei Chang, Prem Natarajan & Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3239–3254. Online: Association for Computational Linguistics.

Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan & Nanyun Peng. 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)*, 3407–3412. Hong Kong, China: Association for Computational Linguistics.

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

Wang, Sida & Christopher Manning. 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)*, 90–94. Jeju Island, Korea: Association for Computational Linguistics.

Webson, Albert, Zhizhong Chen, Carsten Eickhoff & Ellie Pavlick. 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)*, 4090–4105. Online: Association for Computational Linguistics.

Zad, Samira, Joshuan Jimenez & Mark Finlayson. 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)*, 102–113. Online: Association for Computational Linguistics.