How people talk about each other: Modeling Generalized Intergroup Bias and Emotion

Venkata S. Govindarajan[1] **Katherine Atwell**[2] **Barea Sinno**[3]
Malihe Alikhani[2] **David I. Beaver**[1] **Junyi Jessy Li**[1]
[1] Department of Linguistics, The University of Texas at Austin
[2] Department of Computer Sciences, University of Pittsburgh
[3] Department of Political Science, Rutgers University
venkatasg@utexas.edu, kaa139@pitt.edu, barea.sinno@gmail.com,
malihe@pitt.edu, dib@utexas.edu, jessy@utexas.edu

Abstract

Current studies of bias in NLP rely mainly on identifying (unwanted or negative) bias towards a specific demographic group. While this has led to progress recognizing and mitigating negative bias, and having a clear notion of the targeted group is necessary, it is not always practical. In this work we extrapolate to a broader notion of bias, rooted in social science and psychology literature. We move towards predicting interpersonal group relationship (IGR) — modeling the *relationship between the speaker and the target* in an utterance—using fine-grained interpersonal emotions as an anchor. We build and release a dataset of English tweets by US Congress members annotated for interpersonal emotion – the first of its kind, and 'found supervision' for IGR labels; our analyses show that subtle emotional signals are indicative of different biases. While humans can perform better than chance at identifying IGR given an utterance, we show that neural models perform much better; furthermore, a shared encoding between IGR and interpersonal perceived emotion enabled performance gains in both tasks.

1 Introduction

Currently, most work studying bias in NLP situates bias as negative or pejorative language use towards an individual or group based on traits like race, gender, etc (Kaneko and Bollegala, 2019; Sheng et al., 2019; Sap et al., 2020; Webson et al., 2020; Pryzant et al., 2020; Sheng et al., 2020). While these approaches greatly advance our understanding of bias in language and its impact and mitigation in NLP, focusing on specific demographic dimensions or an individual's intent is limiting and not always practical. Research in psychology and social science suggests a different perspective. Bias can be seen as a relationship between people and groups, situated in context (Van Dijk, 2009); as such, bias refers to differences in behavior (in this case language use) as a result of differences in the relationship between speaker and target. The language we produce is biased in one way or another, whether we intend to or not, and whether that bias is positive, negative, or not clearly associated with any valuation (Beaver and Stanley, 2018).

In psychological work on Linguistic Intergroup Bias (Maass, 1999), bias originates from the relationship between the speaker and target of an utterance, i.e. their **interpersonal dynamics**, and manifests later in subtle ways. Consider the utterances (tweets) in (1), drawn from our collected data in which the identity of the speaker and target are masked:

(1) a. **In-group**: We stand w @Doe, who has seen a lot worse than cheap insults from an insecure bully. #MLKDAY weekend.
b. **Out-group**: 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.

Both express support and admiration towards the target referent *Doe* – however, the second example uses words indicative that the speaker and target do not share a relevant social identity (in this case, their political party), expressed by words like *bipartisan*. The intensity of admiration expressed is also greater in (1-a) than (1-b). Thus, these two seemingly similar statements differ along interpersonal dimensions that are instructive as to how the bias of the speaker seeps into the utterance.

We now introduce two new tasks that directly model language use in terms of two interpersonal dimensions: (i) **interpersonal group relationship (IGR) prediction**, where we seek to understand how people talk about others who they consider to be in their same social group (in-group), versus those they consider outside their social group (out-group), and (ii) *perceived* **interpersonal emotion**

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2496–2506

detection, where we situate these differences in terms of the emotion expressed in text *towards or in connection with* a target individual described in the utterance. Note that *interpersonal* emotion is different from a more standard, utterance level emotion detection task, as illustrated in row 2 of Table 1 which has seemingly opposing emotions.

We present a first-of-its-kind, *annotated* dataset for fine-grained interpersonal emotion detection, consisting of 3,033 tweets from members of the US Congress; all of these tweets mention another Congress member, hence providing us with 'found supervision' for IGR prediction (whether the speaker and the target belong to the same political party). Our analyses show that while positive interpersonal emotions appear in both in- and out-group situations, negative emotions like anger and disgust are overwhelmingly present in the latter. Meanwhile, human judgments for in vs. out-group membership on this dataset are overly reliant on the polarity of emotion; specifically, human judges are much *less* likely to attribute positive emotions towards out-group targets.

Baseline performances for perceived interpersonal emotion detection shows that this is a challenging task, as is consistent with existing work in emotion detection in general (Demszky et al., 2020). In particular, emotions in this dataset are often expressed with considerable subtlety, likely a characteristic of official political speech. To investigate whether IGR and emotions are intertwined and useful towards each other, we further developed a multi-task model for the prediction of both. We found compelling evidence that multi-tasking IGR and interpersonal emotion improves performance on both tasks with over 10% improvement in detection of disgust in out-group contexts, and 3% improvement in IGR prediction.

To summarize the contributions of this paper, we tackle *generalized intergroup bias*, a notion of bias rooted in social psychology that applies to all the various differences in the ways that people talk about others in their in-group or out-group. Standard bias tasks in NLP, and the broader goal of debiasing models could thus be set in a more general context. We present the first dataset to study both interpersonal group membership and emotion, which allows us to analyze both human and model behavior in terms of how the two interact with each other. We release our code and data online at `github.com/venkatasg/interpersonal-bias`.

2 Interpersonal Contexts & Emotions

Our aim is to build a generalized, data-driven approach towards studying bias situated in **interpersonal utterances**, which we define as any utterance where there is a target individual being talked about or referred to. Our goal is to model two novel tasks described below; examples are shown in Table 1.

Interpersonal Group relationship IGR is defined by the relationship between the speaker and target of an utterance. People belong to multiple social groups as part of their identity, however usually only some identities are salient in an utterance in context. We define *in-group* utterances as ones where the speaker and target are in the same social group, and *out-group* utterances as one where they are in different social groups. Given an utterance u written by an individual s with target t, the IGR prediction task classifies whether s and t belong to the same social group within the context of u.

Interpersonal Emotion We define *perceived* interpersonal emotion 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. We use the Plutchik wheel of emotions, which is widely adopted in the community, as the basis of our emotion taxionomy (Plutchik, 2001); we use the 8 fundamental emotions (*admiration, anger, disgust, fear, interest, joy, sadness, surprise*) instead of the full 24 emotions in the wheel due to data sparsity. Interpersonal emotions may be different, or a subset of, emotion for the whole of an utterance, as illustrated in rows 2, 3 and 4 of Table 1. Given an utterance u written by an individual s with target t, the interpersonal emotion detection task identifies the perceived emotion of s towards the target t.

3 Data Collection

In our area of focus, we require natural language data which satisfies the following criteria: *(1)* Each utterance must have at least one target about whom the utterance mainly concerns. *(2)* The relationship between the speaker and the target must be inferred based on metadata or other information. Specifically, we are interested in aspects of their social identity that they share or differ on.

The dataset we collect comes from tweets by members of US Congress where other members are mentioned in the same tweet. We use this as

Tweet	Interpersonal Emotion	In/Out group?
As @Doe says, the times have found each and every one of us to Defend our Democracy For The People. Worth reading every line.	Admiration	In-group
Freedom has no greater nor tougher champion than @Doe. My prayers are with him and his family.	Admiration & Sadness	In-group
You don't get to decide what's "fine," @Doe. The constitution does. #DefendOurDemocracy #WednesdayThoughts	Anger & Disgust	Out-group
Thank you again Senator @Doe for leading the SRF WIN Act[...] I'm proud to be a co-sponsor	Admiration & Joy	Out-group

Table 1: Example utterances from our dataset with in/out group and interpersonal emotion labels

a convenient testbed: each member's group affiliation (i.e., their party identity) is public, thus we can easily know whether the speaker is tweeting to a target in their own party or not.[1] In other words, this dataset gives us "found supervision" for our first task of IGR prediction. For our second task, we annotate a subset of these tweets for perceived interpersonal emotion; this is, to our knowledge, the first dataset dedicated to interpersonal emotion.

3.1 Data Sources and Preprocessing

Social media text like tweets offer a fertile ground for our study. A focus on tweets with *mentions* in them satisfies our first criterion – people generally use mentions to say something about or towards another individual on twitter. Tweets by members of US Congress are a matter of public record, and we can infer the social relationship (in terms of party affiliation) between speaker and target using publicly available information. We prioritize working with a dataset of tweets by members of US Congress (downloaded using the Twitter API) between 2010 and 2021, spanning two presidencies, during which both parties held power in Congress. We filter these tweets to exclude retweets, and include those tweets that mention *at most* one other member of Congress whose party affiliation is known. We believe these 2 assumptions are sufficient to arrive at a dataset of tweets where the speaker is talking towards/about *one* target. Thus, we restrict ourselves to two social groups in this sphere — Democrat and Republican parties in the US. We sample an equal number of in-group and out-group tweets from a large sample consisting of all tweets by members of Congress. Apart from years 2010–2012 and 2021 which contained fewer tweets due to sparsity issues, we sampled at least 300 tweets each year.

[1]For simplicity, we do not consider other factors such as the home state of a congress member.

3.2 Interpersonal Emotion Annotation

While we can infer whether a tweet is in-group or out-group based on the identity of speaker and target whose political affiliations are known, we still require annotated data on perceived interpersonal emotions. Interpersonal emotions vary in subtle ways from sentiment or overall sentiment of utterances: an utterance can have negative sentiment overall, but still convey positive emotions towards the target of the sentence (expressing admiration at someone's death for instance). For this reason, we devise an annotation schema for annotating *the emotion expressed by speaker s towards target t.*

Instructions Annotators are presented with a tweet, with the identity of the speaker unknown and that of the target masked with a placeholder name **@Doe** to minimize potential biases of the annotators' prior knowledge of party affiliation intruding into the annotation:

(2) If **@Doe** can get her hair done in person, Congress can vote in person...

Annotators are instructed to read the tweet and select only the most notable emotion(s) they think are expressed by the tweet author *in connection with* **@Doe**. To aid annotators, we provide examples of the 8 Plutchik emotions (*joy, admiration, fear, suprise, sadness, digust, anger and interest*) expressed as interpersonal emotions in tweets. Annotators are also shown a schematic of the Plutchik wheel of emotions, which acquaints them with how the emotions are related to one another in our framework. Annotators are allowed to select more than one emotion to account for emotion co-occurrence. We also explicitly tell annotators that more than one of the emotions can be present in the tweets, to encourage them to select all interpersonal emotions expressed. They are also allowed to not choose any emotion.

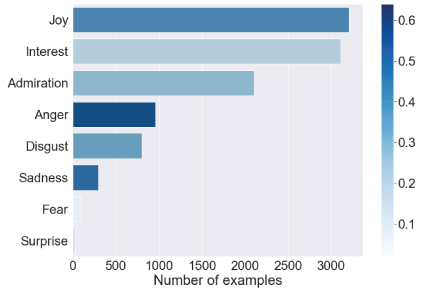

Figure 1: Emotions ordered by the number of examples where at least one rater uses a particular label. The color indicates the average interrater correlation.

Annotation To obtain reliable annotations, we prequalify annotators using a qualifying task. Annotators were recruited on Mechanical Turk using a qualifying task where they were asked to annotate 6 tweets using the schema shown above. We restricted the qualification task to annotators living in the USA who had attempted at least 500 HITS and had a HIT approval rate $\geq 98\%$. After manual inspection, 6 annotators were qualified for bulk annotation. Each tweet was annotated by three different annotators. To ensure annotators were paid a fair wage of at least 10$ an hour, we paid annotators $0.50 per HIT. Each HIT involved annotating 3 tweets, which we estimate to take on average 3 minutes to complete. In total, 3,033 tweets between 2010 and 2021 were annotated with perceived interpersonal emotion.

Agreement To measure agreement between annotators on the Plutchik-8 emotion wheel, we use the Plutchik Emotion Agreement (PEA) score from Desai et al. (2020). The PEA score addresses the issue of penalizing all disagreements equally, by penalizing dissimilar emotion annotations higher than more similar ones (according to the Plutchik wheel). Our PEA score is 0.73. The original PEA formulation used the best(max) pair of emotion annotations between two workers. Taking the *worst* combination of emotions between two workers (averaged over all tweets and workers), the PEA (min) score is 0.60. Overall, we find moderate to high agreement on fine-grained interpersonal emotions. In Figure 1 we also present interrater correlation, a metric used in Demszky et al. (2020); we see that distributions are similar.

Aggregation We consider a tweet to have a certain emotion label if at least 2 out of 3 annotators agree that the particular emotion was present in the tweet. A total of 638 tweets have no interpersonal emotion associated with them. We employ a 80-10-10 train-dev-test split on our data.

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

Table 2: 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
Interest	22.9	27.2	18.6
Joy	26.7	32.2	21.4
Sadness	2.5	2.6	2.4
No Emotion	16.8	14.5	19.1

Table 3: Proportion of emotions in different interpersonal contexts

The number of annotated examples (tweets) per emotion is shown in Table 2. We omit *fear* and *surprise* from future tables due to the absence of annotated examples.

4 Preliminary Analysis

How are emotions distributed? When observing the distribution of aggregated emotion labels themselves, a clear pattern emerges as seen in Table 3. Negative emotions such as anger and disgust are almost always expressed in out-group settings, while positive emotions are present in both in-group and out-group settings. A similar distribution of emotions was observed for Democrats and Republicans — members of both parties reserved their public anger and disgust for members of the other party. This reflects an innate bias in terms of the distribution of interpersonal emotions per situation, and warrants future work to explore negative interpersonal emotions in an in-group setting.

Figure 2 shows the co-occurrence of interpersonal emotions in our dataset. We can see that emotions that are farther apart and more dissimilar, such as admiration and disgust, joy and sadness, co-occur infrequently. Emotions that are closer such as anger and disgust, admiration and joy, co-occur much more often. The only outlier is the higher than normal co-occurrence of admiration with sadness — after a closer examination, this can

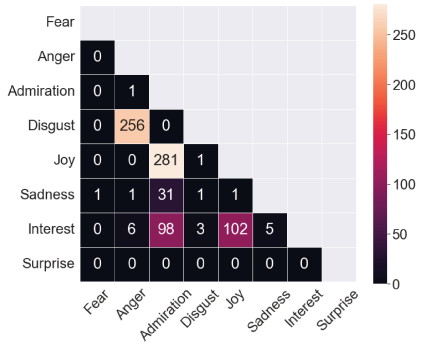

Figure 2: Co-occurence of emotions in our dataset.

be attributed to tweets expressing admiration and sadness at the passing, or end of the career, of a fellow congressperson.

Who were the targets of negative emotions? On further analysis, it appears that most of the out-group disgust and anger is directed at 3 handles – @speakerryan, @speakerpelosi, and @speakerboehner who were all Speakers of the House of Representatives over most of the time period of our dataset. 63.7% of disgust and 64.3% of anger is directed towards these three twitter handles. 11.9% of all tweets in our dataset are directed at these handles, indicating the preponderance of negative interpersonal emotion directed at the Speaker of the house. However, we note that negative emotions like anger and disgust were still expressed towards 51 and 45 different individuals in our dataset, respectively.

Can humans predict in/out-group? While our data naturally comes with "gold" IGR labels, what is unexplored is whether the distinction between in-group and out-group speech is prominent and noticeable by humans. Additionally, it is also unclear if humans might have their own expectation of how in/out-group speech should be characterized.

Concretely, 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, indicating moderate agreement.

To check how accurate their judgements were, we calculate for each annotator their F1 score against our "gold" in/out group labels. Their F1 scores on these 50 tweets were 0.67 and 0.63, which as we will discuss in Section 6, only match simple baselines of supervised systems. Annotators comments indicate that they overly relied on the sentiment of tweets to make the classification — positive sentiment means in-group and negative sentiment means out-group. While negative emotions are over-represented in out-group situations as Table 3 shows, our dataset contains a substantial presence of out-group tweets with positive interpersonal emotions as well. Annotators also noticed some lexical cues like 'bipartisan' that are indicative of out-group tweets.

Do pre-trained representations capture interpersonal emotions? Pre-trained language models have been found to learn sentence representations that cluster by domain without supervision (Aharoni and Goldberg, 2020). We wished to investigate if any of our annotated properties cluster inherently in reduced representations of the tweets in our data. To obtain unsupervised representations, we use BERTweet (Nguyen et al., 2020), a language model pre-trained on 850M English tweets. We take the 768 dimensional embeddings from the final layer of the <s> token in BERTweet, and dimensionally reduce them to 2 dimensions using UMAP (Sainburg et al., 2021). Figure 3 shows the distribution of tweets, color coded for interpersonal emotions. While there is a lot of overlap between representations when stratified by emotion, we can see that some emotions that are intuitively opposite, like admiration & disgust, joy & sadness are moderately separable. This indicates that interpersonal emotions do define some topic or domain level properties of a tweet.

5 Experiments

We detail our experiments for the two novel tasks discussed in Section 2: predicting the IGR (in-group or out-group) given a tweet, and predicting the interpersonal emotion given a tweet. We present baselines for the two tasks separately, and also present a multi-task model to gauge the extent to which knowledge of IGR may help in predicting interpersonal emotion, and vice versa.

5.1 Interpersonal Group Relationship

Sentiment-Rule Our first baseline is a rule-based one leveraging coarse sentiment: if a tweet's sentiment is predicted to be negative, classify it as out-group; if positive, classify it as in-group; and if

Figure 3: Distribution of interpersonal emotions in unsupervised representations of tweets in our dataset. Orange indicates the emotion was present for that tweet. Each point represents one tweet from our dataset.

neutral, classify it as either in-group or out-group randomly. We use a RoBERTa-Base model finetuned for sentiment on tweets (Barbieri et al., 2020) to extract the sentiment of each tweet in our dataset.

NB-SVM As a second baseline, we build an SVM model that uses Naive-Bayes log-counts ratios of unigrams and bigrams (Wang and Manning, 2012).

BERTweet We use BERTweet (Nguyen et al., 2020), a language model pre-trained on 850M English tweets as our dataset consists purely of English language tweets. A classification head is placed on top of the language model. We also experiment with a version where the language model parameters are frozen, and only the classification head parameters are finetuned (BERTweet-ft).

The input to all models is only the tweet with no other context, and the target masked with a placeholder `@USER`.

5.2 Interpersonal Emotion

EmoLex As a baseline model for interpersonal emotion identification, we rely on EmoLex (Mohammad and Turney, 2013). EmoLex consists of 14,182 crowdsourced words associated with the 8 basic Plutchik emotions. Critically, these words appear in emotional contexts, but are not necessarily emotion words themselves. EmoLex counts occurrences of words from its lexicon in an utterance, and assigns a normalized score for each emotion based on occurrence frequency. We consider an emotion to be on, if it's normalized score is $\geq$ 0.001. While EmoLex has issues with regards to its context insensitivity and the social biases built into its lexicon (Zad et al., 2021), we include it as a baseline to understand to what extent interpersonal emotions can be deduced using a lexicon.

BERTweet We use the same BERTweet model as earlier. We add a dense output layer on top of the pretrained model for the purposes of finetuning, with a sigmoid cross entropy loss function to support multi-label classification. The loss is weighted for each of the 8 emotion labels with the ratio of positive and negative examples to increase precision. If none of the 8 emotion labels are flipped on, we consider that to be the 'No Emotion' label, i.e. there is no interpersonal emotion between speaker and target in the tweet. We experiment with a version of the model where the language model parameters are frozen and only the labelling head parameters are finetuned (BERTweet-ft).

5.3 Multi-Task Model

In § 4, we observed that the emotions anger and disgust are overwhelmingly present in out-group situations. Thus, we hypothesize that IGR information would be useful towards interpersonal emotion identification, and vice versa. To test this hypothesis we train a multi-task model. The model is trained to predict *both* the IGR label and emotion using shared parameter finetuning.

We use the same BERTweet model as earlier. We add two dense output layers on top of the pretrained model, one for classifying IGR and another for labelling interpersonal emotion. Both heads share the same parameters below. These are trained with same loss as earlier individual models. The model alternates between finetuning for group relationship and emotion over every training item.

5.4 Implementation

We use `bertweet-base` pretrained embeddings from Huggingface's models hub (Wolf et al., 2020). All models are finetuned for a maximum of 20 epochs with early stopping. Early-stopping patience for models trained on each task separately is 3. The patience for the multi-task model is set at 5 as the multi-tasking setup led to slower convergence. The learning rate for the classification heads was set at 5e-3 while the learning rate for the internal language model parameters was set at 2e-5. Dropout probabilities in classification heads

Model	F1	Model	F1
Majority class	51.1	BERTweet	74.1 (3.3)
Sentiment-Rule	56.3	BERTweet-ft	66.5 (1.6)
NB-SVM	62.5	Multitask	77.3 (0.8)
Human	66.7		

Table 4: Results on test set, with SD in parentheses, for interpersonal group relationship prediction task.

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

Table 5: Top unigram and bigram features from NB-SVM model for each class.

was set at 0.1. The best performing model before early stopping on validation data was chosen in all cases. We report F1 scores averaged over 3 random restarts for all models, with the standard deviation in parentheses next to the mean.

6 Results and Analysis

Interpersonal Group relationship In modeling IGR, we find that `Sentiment-Rule` performs not much better than chance (Table 4). This underscores one strength of our data, which contains a sizable number of out-group tweets with positive interpersonal emotion attached to them. The `NB-SVM` model based on unigrams and bigrams performs slightly better, and picks up on some obvious out-group lexical cues like the lemma 'bipartisan', as shown in Table 5. The `BERTweet` model performs substantially better, performing over 10 points better than humans. The model, with only the classification head finetuned, leaving the language model parameters intact(`BERTweet-ft`) performs about 10 points above chance — indicating that there may be features advantageous towards this task in the vanilla LM itself.

Interpersonal Emotion We find that the `EmoLex` baseline, which relies purely on lexical cues, performs dismally on our data, with poor performance in both in-group and out-group settings(Table 6). This is a strong indication that emotions are expressed more implicitly in this dataset. The `BERTweet` model performs substantially better, indicating that interpersonal emotions, even if implicit, can be learned.

Multitask Model As Table 4 shows, Multitasking the two tasks leads to a noticeable improvement in F1 for IGR prediction, with the differences being statistically significant using a bootstrap test ($p<0.05$; Berg-Kirkpatrick et al., 2012); the multi-task model is also more stable with much lower variance across runs. These results suggest that interpersonal emotion is useful towards IGR prediction.

	Emo Lex	BERTweet	BERTweet -ft	Multi-task
Admir.	37.5	70.3 (3.7)	40.7 (1.1)	68.9 (1.6)
Anger	26.6	71.3 (11.2)	23.0 (3.4)	69.3 (3.3)
Disgust	25.5	47.1 (21.6)	13.0 (4.1)	74.5 (7.1)
Interest	0	53.1 (3.3)	5.8 (2.4)	51.5 (8.5)
Joy	48.4	85.9 (1.9)	71.3 (1.4)	83.6 (1.3)
Sadness	4.3	11.1 (9.6)	0	33.6 (18.5)
No Emotion	22.2	49.1 (1.2)	43.4 (3.8)	71.6(1.2)

Table 6: F1 scores on test set, with SD in parentheses, for interpersonal emotion labelling task.

Emotion	BERTweet	MultiTask
Admiration	77.9 (2.6)	72.8 (3.9)
Anger	71.7 (9.9)	69.4 (3.4)
Disgust	48.2 (22.4)	75.9 (6.5)*

Table 7: F1 scores on test set, SD in parentheses on out-group tweets. * indicates statistical significance ($p<0.05$)

Table 6 shows that the performance of the multitask model on predicting interpersonal emotions is significantly better that the BERTweet model ($p<0.05$) on emotions like *disgust*, which suggests that IGR is useful towards the task of emotion identification. Furthermore, multitasking boosted performance at predicting the *no emotion* label by 20%. Table 7 compares the multitask model's performance against the BERTweet model in *out-group* settings (where most of the gains were found) for 3 emotions — illustrating the boost in performance afforded by joint modeling of IGR and emotion for *digust*. The 3 emotions listed also showed significant differences in their distribution in in-group and out-group settings.

Humans vs. Models Comparatively, we find that model performance exceeds human performance on the task of in-group versus out-group prediction, albeit not on the same dataset. The model's main driver of performance is its high accuracy on positive intergroup emotion out-group tweets, such as those expressing admiration or joy. Human annotators consistently fall back on the heuristic that sentences with positive affect probably imply that

the speaker is talking about someone in their in-group. But it is not the case in the political domain, where overtures to bipartisanship serve as useful signals. For instance, both (3-a) and (3-b) express admiration towards the target Doe, where the first is in-group while the second is out-group. The call to civility is the only subtle linguistic cue that this tweet may constitute out-group speech.

(3) a. Admire @OfficialCBC Chairman @Doe's moral voice on issues of racism and restorative justice. He is a real leader for our nation and Congress.

b. A decade has passed, but our friendship is the same. Proud to work with @Doe to #ReviveCivility. #tbt Read more about our efforts here:

Future work needs to look into what information the embeddings are using to make their classification decision.

Model Errors While the multitasking setup improves model performance on the task of predicting IGR, and outperforms human labelers in our small pilot, it still gets some easy examples wrong, such as labelling (4) as in-group even though it expresses some disgust at the target. The model also falls into the same trap as human labelers — for instance assuming that a tweet expressing admiration must be in-group (5).

(4) Trump selected @USER for HHS Secretary. Price has undeniable history of cutting access to healthcare to millions, especially women.

(5) Inspiring speech from @USER - we have a duty to represent our country with respect & dignity. #NationalDayofCivility.

To ensure that model performance on IGR prediction is not limited by the size of our training data, we experimented with training `BERTweet` models on larger datasets. Since we have 'found supervision' for IGR labels, we only need to increase training data size by sampling more tweets from relevant accounts using the same procedure detailed in § 3.1. We found that F1 score does not increase with more training data.

Future work needs to look into linguistically motivated ways to improve model performance on the IGR task. Since we have observed that the multi-task setup improves model performance, perhaps other multi-task setups, such as modeling the overall affect towards the target expressed by the speaker might help in modeling IGR better.

7 Related Work

Emotion and Stance Detection A wealth of work has looked at corpora and models for the detection of perceived emotion in social media text (Mohammad, 2012; Wang et al., 2012; Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017; Desai et al., 2020; Demszky et al., 2020). However existing work doesn't distinguish between emotion of a sentence as a whole, versus interpersonal emotion towards a target. The task closest to our study of interpersonal emotions is stance detection: whether the author has a favourable, neutral, or negative position towards a proposition or target. Mohammad et al. (2016) looked at stance in five target domains are given: abortion, atheism, climate change, feminism and Hillary Clinton. While stance detection focuses on a collection of utterances with the same topic, our interest is in modeling interpersonal emotion towards a target individual which is more fine-grained and can vary in each utterance.

Intergroup bias in Psychology The Linguistic Intergroup Bias (LIB) theory (Maass et al., 1989; Maass, 1999) states that there is a systematic asymmetry in language production qualities of a speaker as a function of the social category to which the referent of an utterance belongs. Through psycholinguistic experiments, LIB seeks to explain why stereotypes are transmitted and persist in daily life: in an interpersonal situation, socially desirable in-group behaviors and undesirable out-group behaviors are encoded at a higher level of **abstraction**, whereas socially undesirable in-group behaviors and desirable in-group behaviors are encoded at a lower level of abstraction. Work in psychology and psycholinguistics reproduced LIB in various domains such as political news reporting (Anolli et al., 2006) and crime reporting (Gorham, 2006); as well as work exploring how LIB can be used as an indicator for a speaker's prejudicial attitudes (Hippel et al., 1997), or as a predictor for racism (Schnake and Ruscher, 1998).

Contemporaneous studies on LIB, however, are hand-coded and have so far tended to focus on narrow concepts such as abstractness of the verb and coarse notions of sentiment. Nonetheless, the LIB hypothesis connects the two dimensions of interpersonal dynamics studied here with a third dimension directly related to semantic properties of the utterance.

8 Conclusion

Taking a cue from studies of bias in social science and psychology, we situate bias in language use through the lens of interpersonal relationships between the speaker and target of an utterance, and the speaker's interpersonal emotional state with respect to the target. Over a corpus of tweets by members of US Congress, we introduce two novel tasks – interpersonal group relationship prediction (IGR) and interpersonal emotion labelling, to better understand variation in language as a function of social relationship between speaker and target in interpersonal utterances. We find certain interpersonal emotions like anger and disgust are overrepresented in out-group situations, with the majority of the negative emotions directed at leaders of the two political parties. Through modeling studies, we find that transformer based models perform better than humans at predicting IGR given an utterance, raising the question as to what latent features of language the model uses to make this decision. Finally, we also find that joint modelling of the two dimensions is beneficial to prediction of certain interpersonal emotions in out-group situations. Future work needs to look into what information is useful for predicting IGR and emotions – with the Linguistic Intergroup Bias literature offering a clue as to which higher level semantic features vary systematically.

Ethics Statement

For our corpus of tweets on which we performed annotations, we downloaded the tweets using the official Twitter API. In accordance with the Twitter Terms of Service, we release tweet IDs and usernames, but not the tweet text itself. Our dataset was built through crowdsourced annotations on Amazon Mechanical Turk. To ensure annotators were paid a fair wage of at least $10 an hour, we paid annotators $0.50 per HIT. Each HIT involved annotating 3 tweets, which we estimate to take on average 3 minutes to complete.

Limitations

Our results show the importance of having reliable and accurate emotion prediction models, which is an open problem in psychology and computer science. Future work might look into identifying different fine-grained emotional constructs and study their correlations with the underlying linguistic biases. Future work may also look into the generalizability of the results presented here in other domains of language use.

While we present the utterances as constituting natural speech by one speaker (the congressperson who sent the tweet), it is likely most congresspeople employ social media teams that help in crafting the language of some of their tweets. However, we believe for the sake of interpersonal group membership, the relationship between the speakers and their targets would not be affected.

Finally, while we show that transformer based models perform better at IGR prediction than humans, we note that the human performance was on a small subset of test data. While it is possible that these models discovered latent features that could explain their better performance, the model could also be using spurious features idiosyncratic to our dataset, rather than true differences in in-group versus out-group speech.

Acknowledgements

This research is partially supported by NSF grants IIS-2107524, IIS-2145479 and Good Systems,[2] a UT Austin Grand Challenge to develop responsible AI technologies. We also acknowledge the Texas Advanced Computing Center (TACC)[3] at UT Austin for providing the computational resources for many of the results within this paper.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Luigi Anolli, Valentino Zurloni, and Giuseppe Riva. 2006. Linguistic Intergroup Bias in Political Communication. *The Journal of General Psychology*, 133:237 – 255.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval:

[2] http://goodsystems.utexas.edu

[3] https://www.tacc.utexas.edu

Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

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

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

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

W. Hippel, Denise Sekaquaptewa, and P. Vargas. 1997. The Linguistic Intergroup Bias As an Implicit Indicator of Prejudice. *Journal of Experimental Social Psychology*, 33:490–509.

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

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

Anne Maass, Daniel Anthony Salvi, Luciano Arcuri, and Gün R. Semin. 1989. Language use in intergroup contexts: the linguistic intergroup bias. *Journal of Personality and Social Psychology*, 57 6:981–93.

Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, 31:301 – 326.

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

Dat Quoc Nguyen, Thanh Vu, and 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*, pages 9–14. Association for Computational Linguistics.

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

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

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and 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*, pages 5477–5490, Online. Association for Computational Linguistics.

Sherry B Schnake and Janet B Ruscher. 1998. Modern Racism as a predictor of the Linguistic Intergroup Bias. *Journal of Language and Social Psychology*, 17(4):484–491.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and 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)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

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

Sida Wang and 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)*, pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing Twitter "Big Data" for Automatic Emotion Identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and 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)*, pages 4090–4105, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Samira Zad, Joshuan Jimenez, and 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)*, pages 102–113, Online. Association for Computational Linguistics.