Dimensions of Interpersonal Dynamics in Text: Group Membership and Fine-grained Interpersonal Emotion

Venkata S. Govindarajan1 Katherine Atwell2 Barea Sinno3 Malihe Alikhani2 David I. Beaver1 Junyi Jessy Li1
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

Abstract

The ability of language to perpetuate inequality is most evident when individuals refer to, or talk about, other individuals in their utterances. While current studies of bias in NLP rely mainly on identifying hate speech or bias towards a specific group, we believe we can reach a more subtle and nuanced understanding of the interaction between bias and language use by modeling the speaker, the text, and the target in the text. In this paper, we introduce a dataset of 3033 English tweets by US Congress members annotated for interpersonal emotion, and ‘found supervision’ for interpersonal group membership labels. We find that negative emotions such as anger and disgust are used predominantly in out-group situations, and directed predominantly at leaders of opposite parties. While humans can perform better than chance at identifying interpersonal group membership given an utterance, neural models perform much better; furthermore, a shared encoding between interpersonal group membership and interpersonal perceived emotion enabled some performance gains in the latter. This work aims to re-align the study of bias in NLP away from specific instances of bias to one which encapsulates the relationship between speaker, text, target and social dynamics.

1 Introduction

Language has the power to perpetuate and reinforce social biases and stereotypes. In addition to overt displays of bias, dogwhistles, ironic humour, dark jokes, sarcasm and other forms of non-ideal language (Beaver and Stanley, 2018) play a big role in perpetuating bias covertly. The existence of bias in natural language data has led to recent research investigating bias in socio-political contexts (Sap et al., 2020), and to de-bias datasets and models, by e.g., manipulating the word embedding space (Kaneko and Bollegala, 2019; Webson et al., 2020), and controlled text generation (Pryzant et al., 2020; Sheng et al., 2020).

While these approaches greatly advance our understanding of language bias and its impact and mitigation in NLP, they rely on a presumption of communicative intent to express bias. As work in sociology and psychology has shown (Maass, 1999), expressions of bias do not necessarily involve communicative intent; bias originates from the relationship between the speaker and target of an utterance, i.e. their interpersonal dynamics, and manifests later in usually subtle ways. There are several dimensions of interpersonal dynamics which influence language use like power (Hancock et al., 2010) and spatial distance (Rashid and Blanco, 2017), but for this paper we wish to focus on two dimensions that are salient and contextualized in the sentiment the speaker expresses towards the target of an utterance, and the speaker and target’s social relationship. Consider the sentences in (1) (from our collected data), where 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. #MLK-DAY weekend.

(1) 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. Thank you Dr. Wood for being a pioneering voice.

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 two interpersonal dimensions that are instructive as to how the bias of the speaker seeps into the utterance.
Following previous work examining how dimensions of interpersonal relationships influence language use, we introduce two new dimensions that directly models language use in terms of the social relationship of the speaker and the target (person/entity under discussion), and the perceived emotional relationship between them. This entails two novel tasks: (i) interpersonal group membership prediction, where we seek to understand how people talk about others who they consider in their same social group (in-group), versus those they consider outside their social group (out-group). (ii) perceived interpersonal emotion classification, where we situate these differences in terms of the emotion expressed in text towards or in connection with the target.

For the first task, we collect a dataset and model interpersonal group memberships on English language tweets in the political domain. This domain allows us to get “found supervision”, using a politician’s party as in-out group labels, and explore how language varies covertly in tweets where the speaker and the target share political party identity, versus when they don’t. Specifically, we collect a new dataset of interpersonal group memberships and emotions on social media tweets from US Congresspeople between 2010 and 2021. For the second task, we sample 3033 tweets for the annotation of perceived, fine-grained interpersonal emotions. Importantly, we mask the identity of the speaker and the target such that the annotators are blind to the party identity of either (thus in-out group labels are unknown), and are not biased when providing labels. To our knowledge this is the first annotated dataset in computational linguistics dedicated to the detection of interpersonal emotions.

Our analyses show that while positive interpersonal emotions appear in both in- and out-group situations, certain emotions like anger and disgust are overwhelmingly present in the latter. Meanwhile, human judgments for in-out group on this dataset are overly reliant on the polarity of emotion; namely, people are much less likely to attribute positive emotions to out-group targets. This surprising “bias” reflects to a much lesser extent on models.

Baseline performances for perceived interpersonal emotion classification 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 expressed very implicitly, likely a characteristic of official political speech. To investigate whether interpersonal group membership and emotions are intertwined and useful towards each other, we further design a multi-task model for the prediction of both. We found compelling evidence that group membership informs interpersonal emotion prediction in out-group situations with over 10% improvement in detection of admiration and disgust, although not vice versa. We release our code and data at https://github.com/venkatasg/Interpersonal-Dynamics.

2 Interpersonal Contexts & Emotions

Currently, most work studying bias in NLP situates bias as negative or pejorative language use towards an individual or group (Sap et al., 2020; Kaneko and Bollegala, 2019; Webson et al., 2020). This narrow focus of bias requires explicit communicative intent on the part of the speaker. In contrast, we take a broader view following research in psychology and social science and see bias as a relationship between people and groups, situated in context (Van Dijk, 2009); as such, bias means a change in behavior (in this case language) as a result of a change in relationship between speaker and target. With this in mind, we aim for a generalized, data-driven approach 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 Membership Interpersonal group membership 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 interpersonal group membership 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 speaker \( s \) towards, or in connection with the target
As @speakerpelosi says, the times have found each and every one of us to Defend our Democracy For The People. Worth reading every line.

Freedom has no greater nor tougher champion than @senjohnmccain. My prayers are with him and his family.

You don’t get to decide what’s “fine,” @lindseygrahamsc. The constitution does. #DefendOurDemocracy #WednesdayThoughts

Thank you again Senator @johnboozman for leading the SRF WIN Act[...]. I’m proud to be a co-sponsor

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

| Tweet                                                                 | Interpersonal Emotion | In/Out group? |
|-----------------------------------------------------------------------|-----------------------|---------------|
| As @speakerpelosi 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 @senjohnmccain. My prayers are with him and his family. | Admiration & Sadness   | In-group      |
| You don’t get to decide what’s “fine,” @lindseygrahamsc. The constitution does. #DefendOurDemocracy #WednesdayThoughts | Anger & Disgust        | Out-group     |
| Thank you again Senator @johnboozman for leading the SRF WIN Act[...]. I’m proud to be a co-sponsor | Admiration & Joy       | Out-group     |

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 a convenient testbed: each member’s group affiliation (i.e., their party identity) is public. In other words, this dataset gives us “found supervision” for our first task of interpersonal group membership 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 offers 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 easily using publicly available information. We prioritize working with a dataset of tweets by members of the US Congress downloaded using the Twitter API between 2010 and 2021 (this time period spanned over two presidencies, and 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 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.

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 $t$. 

---

2For simplicity, we do not consider other factors such as the home state of a congress member.
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. Further, if @JoeBiden can vote in person, Americans should be encouraged to cast their 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, surprise, sadness, disgust, 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. Finally, they are also allowed to not choose any emotion.

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.

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 set-
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

nings, 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 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.

Can humans predict in/out-group? While our data naturally comes with “gold” interpersonal group membership labels, what is unexplored is 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 interpersonal group membership 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 $\kappa$ 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 interpersonal group membership (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 interpersonal group membership may help in predicting interpersonal emotion, and vice versa.

5.1 Interpersonal Group Membership

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 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 Naïve-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
Table 4: Results on test set for interpersonal group membership prediction task.

| Model                | F1   | Model     | F1     |
|----------------------|------|-----------|--------|
| Majority class       | 51.1 | BERTweet  | 77.6(2.3) |
| Sentiment-Rule       | 56.3 | BERTweet-ft | 52.1(1.8) |
| NB-SVM               | 62.5 | Joint     | 76.9(0.5) |
| Human                | 66.7 |           |        |

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

|          | Emo Lex | BERTweet | BERTweet-ft | Joint |
|----------|---------|----------|-------------|-------|
|         |         |          |             |       |
|         | 37.5    | 66.3(3.8)| 37.2(5.8)   | 67.4(0.3) |
|         | 26.6    | 71.4(10.6)| 9.4(8.4)   | 76.0(2.3) |
|         | 25.5    | 57.1(16.7)| 4.3(7.5)   | 80.9(2.1) |
|         | 0       | 52.1(6.9) | 15.9(10.3) | 53.7(7.9) |
|         | 48.4    | 82.7(1.9) | 75.1(3.9)  | 80.0(9.0) |
|         | 4.3     | 37.8(35.6)| 0          | 69.1(9.6) |
|         | 22.2    | 49.4(0.7) | 40.8(1.3)  | 51.2(0.4) |

Table 6: F1 scores on test set for interpersonal emotion labelling task.

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 interpersonal group membership 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 interpersonal group membership 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 pre-trained model, one for classifying interpersonal group membership 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 membership 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 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 Membership In modeling interpersonal group membership, 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 close to chance.

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 Multi-tasking the two tasks does not lead to improvements in F1 for interpersonal context. However, the differences are not statistically significant using a bootstrap test (Berg-Kirkpatrick et al., 2012); the multi-task model is also more stable with much lower variance across
Table 7: F1 scores on test set on out-group tweets. * indicates statistical significance (p<0.05)

| Emotion   | BERTweet | MultiTask |
|-----------|----------|-----------|
| Admiration| 63.8(3.4)| 76.5(4.1)*|
| Anger     | 71.1(10.3)| 75.8(0.5)|
| Disgust   | 58.0(16.9)| 81.3(2.2)*|

The performance of the multitask model is significantly better (p<0.05) on some out-group emotions, which suggests that interpersonal group membership is useful towards the task of emotion identification. Table 7 compares the multitask model’s performance against the BERTweet model in out-group settings — illustrating the boost in performance afforded by joint modeling of interpersonal group membership and emotion for 3 emotions — admiration, anger and disgust. These 3 emotions also showed significant distributions in their proportion 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.

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 try to 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 membership prediction 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 over-represented in out-group situations, with 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 interpersonal group membership 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 interpersonal group membership 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.

Acknowledgements

This research is partially supported by NSF grants IIS-2107524, IIS-2145479 and Good Systems, a UT Austin Grand Challenge to develop responsible AI technologies. We also acknowledge the Texas Advanced Computing Center (TACC) 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: 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 Sujiith 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.

Jeffrey T. Hancock, David I. Beaver, Cindy K. Chung, Joey Frazee, James W. Pennebaker, Art Graesser, and Zhiquiang Cai. 2010. Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *Behavioral Sciences of Terrorism and Political Aggression*, 2(2):108–132.

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. Crowd-sourcing 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: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadow 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.

Farzana Rashid and Eduardo Blanco. 2017. Dimensions of interpersonal relationships: Corpus and experiments. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2307–2316, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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 Conference 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.