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

TEXAS
The University of Texas at Austin

Venkata S Govindarajan*, Katherine Atwell[†], Barea Sinno[‡], Malihe Alikhani[†], David I. Beaver*, Junyi Jessy Li*
*The University of Texas at Austin, [†]University of Pittsburgh, [‡] Rutgers University

INTERPERSONAL BIAS

Bias can be analyzed as behavioral differences situated in varied social relationships. Consider these 2 utterances (tweets):

(1) We stand w @Doe, who has seen a lot worse than cheap insults from an insecure bully. #MLKDAY weekend.

(2) 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.

Words like *bipartisan* in (2) delicately signal that the speaker and target (@Doe) do not belong to the same social group; furthermore the emotion (admiration) expressed in (1) is more effusive. We use these insights to study 2 dimensions of interpersonal bias — intergroup relationship & interpersonal emotion.

DATA & ANALYSIS

We build a dataset of 3033 interpersonal tweets by members of U.S. Congress with the following properties:

- *found supervision* for intergroup relationship (in-group and out-group)
- annotations for fine-grained interpersonal emotions based on the Plutchik wheel, blind to speaker and target.

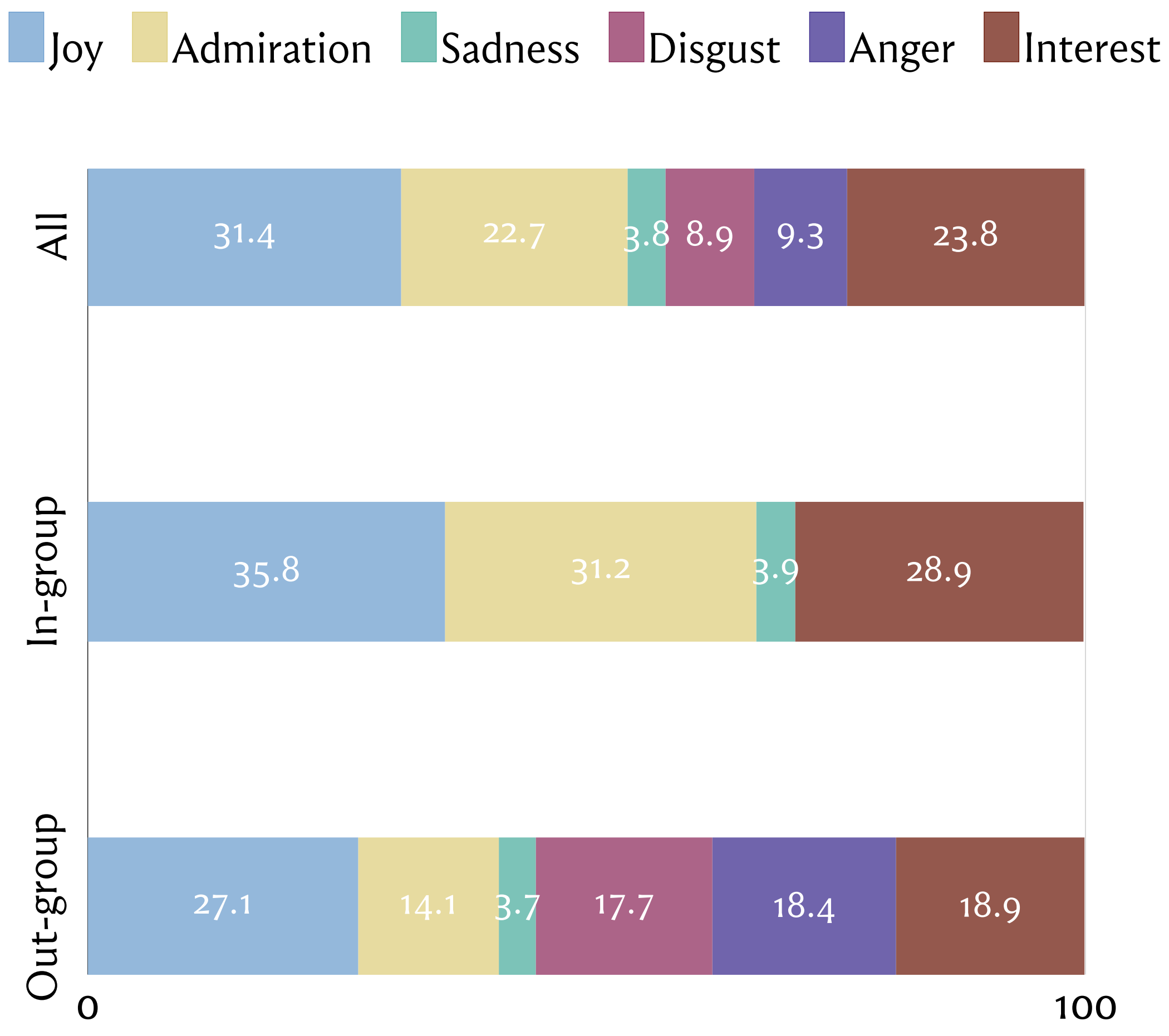

Figure 1: Distribution of interpersonal emotions across contexts.

As Figure 1 shows, negative emotions are overwhelmingly present in the out-group context. Most disgust and anger is directed at the out-group, and generally at 3 users — the party leaders at the time.

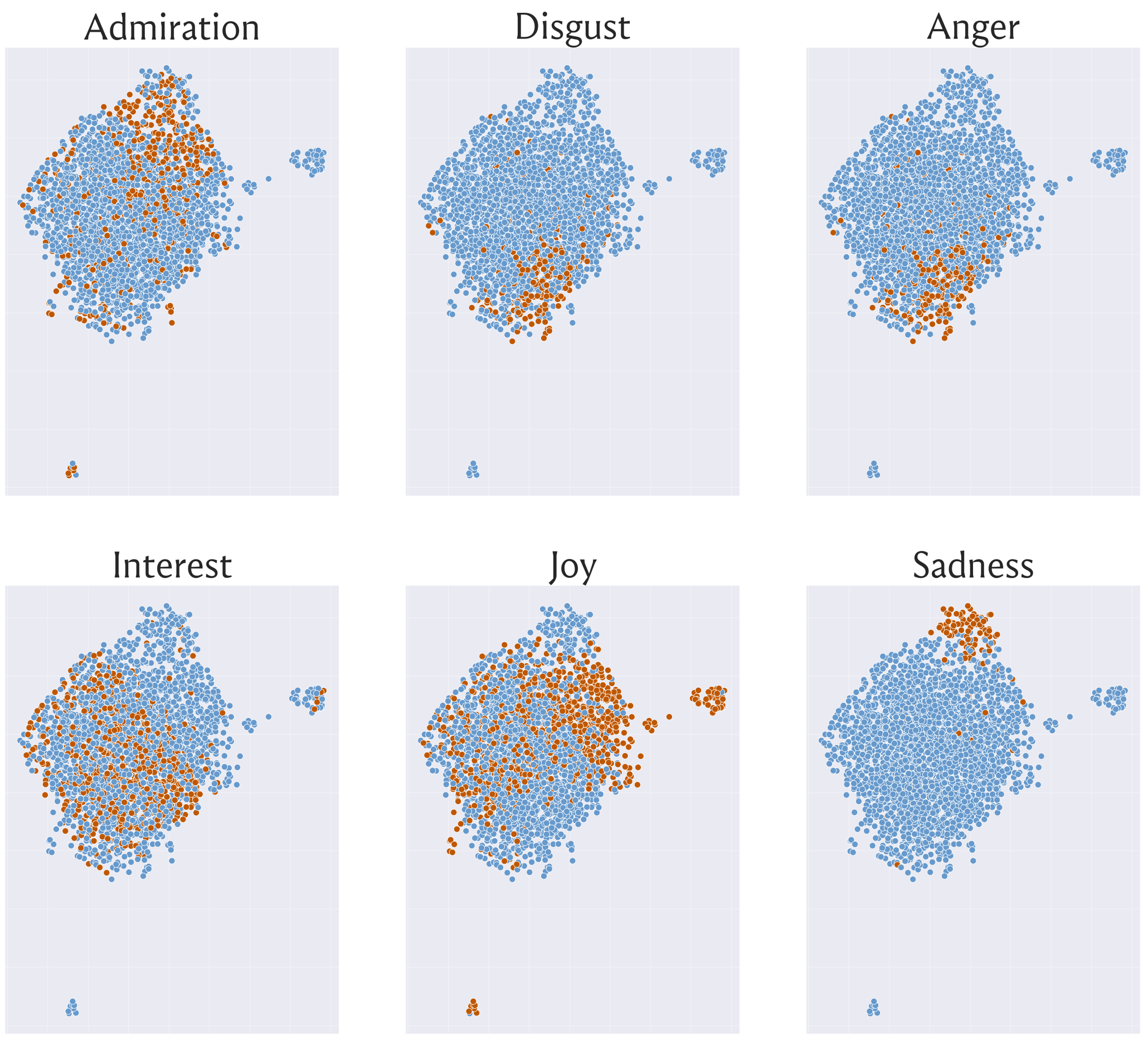

Figure 2: Distribution of emotions in unsupervised representations of tweets. Orange indicates the emotion was present.

Observing unsupervised representations (after dimension reduction) of the tweets in Figure 2, emotions that are intuitively opposite, like admiration & disgust, joy & sadness, are moderately separable. This indicates that interpersonal emotions capture some topic or domain level properties of a tweet.

KEY CONTRIBUTIONS

- Taking a cue from bias research in social science and psychology (Maass 1999), 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.
- A dataset of 3033 English tweets by members of U.S. Congress annotated for interpersonal emotion, with *found supervision* for intergroup relationship (in-group and out-group).
- Interpersonal emotion and intergroup relationship systematically interact, as evidence in Figure 2, and further fortified by multitasking results in Figures 3 and 4.

Check out our code and data online at: `github.com/venkatasg/interpersonal-bias`

MODELING

We want to model and predict Intergroup Relationship and Interpersonal Emotion to delve deeper into their interaction, and to study model behavior in the future.

Baseline Predict each dimension separately using lexical features.

BERTweet Predict each dimension separately by finetuning an LM.

Multitask Train and predict both dimensions *jointly*, using a shared LM encoding with only separate classifier layers.

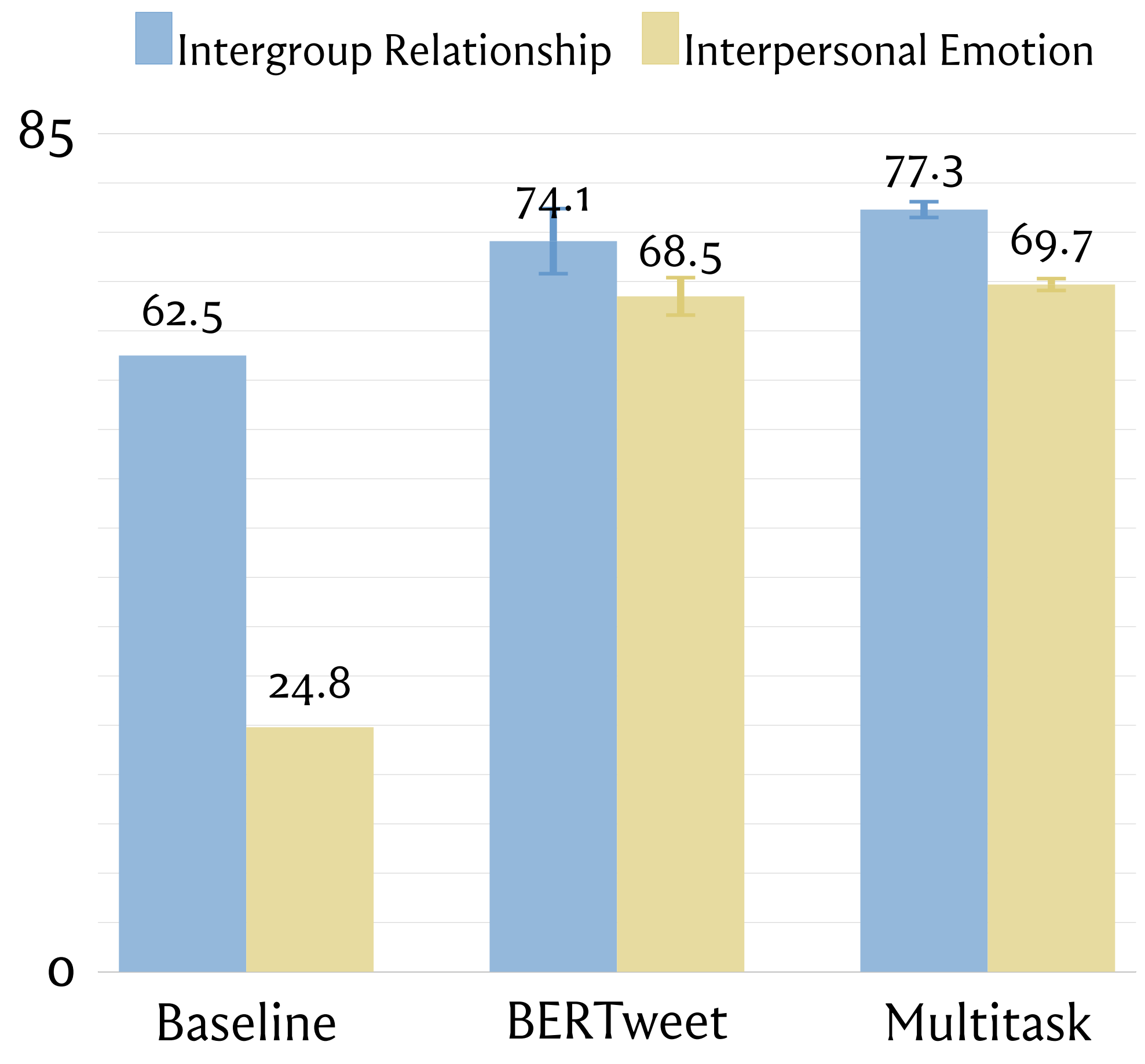

Figure 3: F1 scores on 2 tasks for 3 different models.

Figure 3 shows that the multitask model performs best — interpersonal emotion does help in intergroup relationship prediction, reinforcing our earlier observation.

Intergroup relationship helps in predicting certain interpersonal emotions like disgust, sadness & 'No emotion' as shown in Figure 4, with most of the gains from multitasking coming in out-group settings.

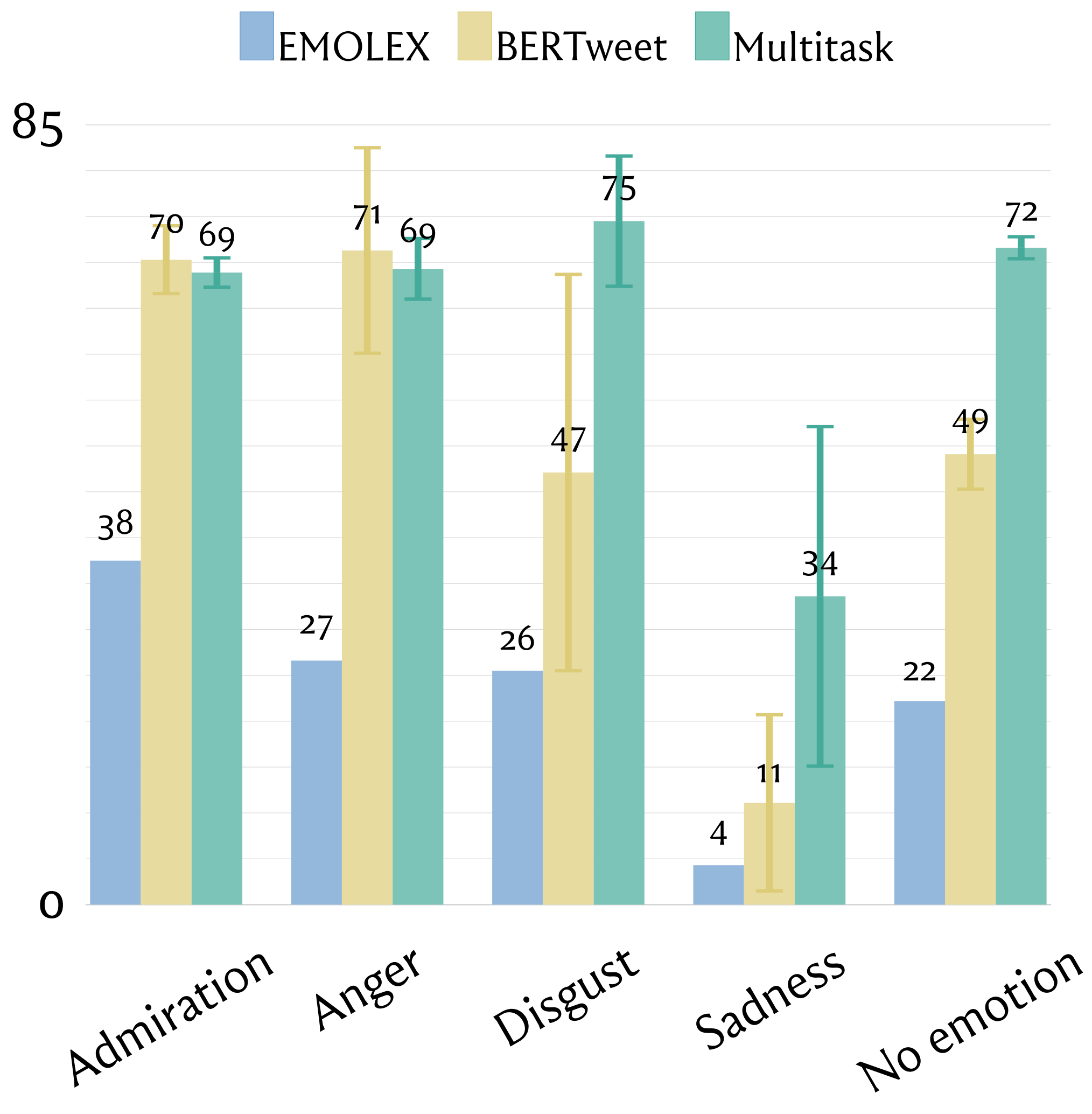

Figure 4: F1 scores on each emotion for all models.

FUTURE WORK

- Which linguistic features explain the systematic variation between in-group and out-group language?
- How generalizable are the results to other domains with more situated utterances?