GoEmotions: A Dataset of Fine-Grained Emotions

Dorottya Demszky[1]* **Dana Movshovitz-Attias**[2] **Jeongwoo Ko**[2]
Alan Cowen[2] **Gaurav Nemade**[2] **Sujith Ravi**[3]*
[1]Stanford Linguistics [2]Google Research [3]Amazon Alexa
ddemszky@stanford.edu
{danama, jko, acowen, gnemade}@google.com
sravi@sravi.org

Abstract

Understanding emotion expressed in language has a wide range of applications, from building empathetic chatbots to detecting harmful online behavior. Advancement in this area can be improved using large-scale datasets with a fine-grained typology, adaptable to multiple downstream tasks. We introduce GoEmotions, the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We demonstrate the high quality of the annotations via Principal Preserved Component Analysis. We conduct transfer learning experiments with existing emotion benchmarks to show that our dataset generalizes well to other domains and different emotion taxonomies. Our BERT-based model achieves an average F1-score of .46 across our proposed taxonomy, leaving much room for improvement.[1]

Sample Text	Label(s)
OMG, yep!!! That is the final answer. Thank you so much!	gratitude, approval
I'm not even sure what it is, why do people hate it	confusion
Guilty of doing this tbph	remorse
This caught me off guard for real. I'm actually off my bed laughing	surprise, amusement
I tried to send this to a friend but [NAME] knocked it away.	disappointment

Table 1: Example annotations from our dataset.

1 Introduction

Emotion expression and detection are central to the human experience and social interaction. With as many as a handful of words we are able to express a wide variety of subtle and complex emotions, and it has thus been a long-term goal to enable machines to understand affect and emotion (Picard, 1997).

In the past decade, NLP researchers made available several datasets for language-based emotion classification for a variety of domains and applications, including for news headlines (Strapparava and Mihalcea, 2007), tweets (CrowdFlower, 2016; Mohammad et al., 2018), and narrative sequences (Liu et al., 2019), to name just a few. However, existing available datasets are (1) mostly small, containing up to several thousand instances, and (2) cover a limited emotion taxonomy, with coarse classification into Ekman (Ekman, 1992b) or Plutchik (Plutchik, 1980) emotions.

Recently, Bostan and Klinger (2018) have aggregated 14 popular emotion classification corpora under a unified framework that allows direct comparison of the existing resources. Importantly, their analysis suggests annotation quality gaps in the largest manually annotated emotion classification dataset, CrowdFlower (2016), containing 40K tweets labeled for one of 13 emotions. While their work enables such comparative evaluations, it highlights the need for a large-scale, consistently labeled emotion dataset over a fine-grained taxonomy, with demonstrated high-quality annotations.

To this end, we compiled GoEmotions, the largest human annotated dataset of 58k carefully selected Reddit comments, labeled for 27 emotion categories or Neutral, with comments extracted from popular English subreddits. Table 1 shows an illustrative sample of our collected data. We design our emotion taxonomy considering related work in psychology and coverage in our data. In contrast to Ekman's taxonomy, which includes only one positive emotion (joy), our taxonomy includes a large number of positive, negative, and ambiguous emotion categories, making it suitable for downstream

* Work done while at Google Research.

[1]Data and code available at https://github.com/google-research/google-research/tree/master/goemotions.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4040–4054

conversation understanding tasks that require a subtle understanding of emotion expression, such as the analysis of customer feedback or the enhancement of chatbots.

We include a thorough analysis of the annotated data and the quality of the annotations. Via Principal Preserved Component Analysis (Cowen et al., 2019b), we show a strong support for reliable dissociation among all 27 emotion categories, indicating the suitability of our annotations for building an emotion classification model.

We perform hierarchical clustering on the emotion judgments, finding that emotions related in intensity cluster together closely and that the top-level clusters correspond to sentiment categories. These relations among emotions allow for their potential grouping into higher-level categories, if desired for a downstream task.

We provide a strong baseline for modeling fine-grained emotion classification over GoEmotions. By fine-tuning a BERT-base model (Devlin et al., 2019), we achieve an average F1-score of .46 over our taxonomy, .64 over an Ekman-style grouping into six coarse categories and .69 over a sentiment grouping. These results leave much room for improvement, showcasing this task is not yet fully addressed by current state-of-the-art NLU models.

We conduct transfer learning experiments with existing emotion benchmarks to show that our data can generalize to different taxonomies and domains, such as tweets and personal narratives. Our experiments demonstrate that given limited resources to label additional emotion classification data for specialized domains, our data can provide baseline emotion understanding and contribute to increasing model accuracy for the target domain.

2 Related Work

2.1 Emotion Datasets

Ever since Affective Text (Strapparava and Mihalcea, 2007), the first benchmark for emotion recognition was introduced, the field has seen several emotion datasets that vary in size, domain and taxonomy (cf. Bostan and Klinger, 2018). The majority of emotion datasets are constructed manually, but tend to be relatively small. The largest manually labeled dataset is CrowdFlower (2016), with 39k labeled examples, which were found by Bostan and Klinger (2018) to be noisy in comparison with other emotion datasets. Other datasets are automatically weakly-labeled, based on emotion-related hashtags on Twitter (Wang et al., 2012; Abdul-Mageed and Ungar, 2017). We build our dataset manually, making it the largest human annotated dataset, with multiple annotations per example for quality assurance.

Several existing datasets come from the domain of Twitter, given its informal language and expressive content, such as emojis and hashtags. Other datasets annotate news headlines (Strapparava and Mihalcea, 2007), dialogs (Li et al., 2017), fairytales (Alm et al., 2005), movie subtitles (Öhman et al., 2018), sentences based on FrameNet (Ghazi et al., 2015), or self-reported experiences (Scherer and Wallbott, 1994) among other domains. We are the first to build on Reddit comments for emotion prediction.

2.2 Emotion Taxonomy

One of the main aspects distinguishing our dataset is its emotion taxonomy. The vast majority of existing datasets contain annotations for minor variations of the 6 basic emotion categories (joy, anger, fear, sadness, disgust, and surprise) proposed by Ekman (1992a) and/or along affective dimensions (valence and arousal) that underpin the circumplex model of affect (Russell, 2003; Buechel and Hahn, 2017).

Recent advances in psychology have offered new conceptual and methodological approaches to capturing the more complex "semantic space" of emotion (Cowen et al., 2019a) by studying the distribution of emotion responses to a diverse array of stimuli via computational techniques. Studies guided by these principles have identified 27 distinct varieties of emotional experience conveyed by short videos (Cowen and Keltner, 2017), 13 by music (Cowen et al., in press), 28 by facial expression (Cowen and Keltner, 2019), 12 by speech prosody (Cowen et al., 2019b), and 24 by nonverbal vocalization (Cowen et al., 2018). In this work, we build on these methods and findings to devise our granular taxonomy for text-based emotion recognition and study the dimensionality of language-based emotion space.

2.3 Emotion Classification Models

Both feature-based and neural models have been used to build automatic emotion classification models. Feature-based models often make use of hand-built lexicons, such as the Valence Arousal Dominance Lexicon (Mohammad, 2018). Using representations from BERT (Devlin et al., 2019), a

transformer-based model with language model pre-training, has recently shown to reach state-of-the-art performance on several NLP tasks, also including emotion prediction: the top-performing models in the EmotionX Challenge (Hsu and Ku, 2018) all employed a pre-trained BERT model. We also use the BERT model in our experiments and we find that it outperforms our biLSTM model.

3 GoEmotions

Our dataset is composed of 58K Reddit comments, labeled for one or more of 27 emotion(s) or Neutral.

3.1 Selecting & Curating Reddit comments

We use a Reddit data dump originating in the reddit-data-tools project[2], which contains comments from 2005 (the start of Reddit) to January 2019. We select subreddits with at least 10k comments and remove deleted and non-English comments.

Reddit is known for a demographic bias leaning towards young male users (Duggan and Smith, 2013), which is not reflective of a globally diverse population. The platform also introduces a skew towards toxic, offensive language (Mohan et al., 2017). Thus, Reddit content has been used to study depression (Pirina and Çöltekin, 2018), microaggressions (Breitfeller et al., 2019), and Yanardag and Rahwan (2018) have shown the effect of using biased Reddit data by training a "psychopath" bot. To address these concerns, and enable building broadly representative emotion models using GoEmotions, we take a series of data curation measures to ensure our data does not reinforce general, nor emotion-specific, language biases.

We identify harmful comments using pre-defined lists containing offensive/adult, vulgar (mildly offensive profanity), identity, and religion terms (included as supplementary material). These are used for data filtering and masking, as described below. Lists were internally compiled and we believe they are comprehensive and widely useful for dataset curation, however, they may not be complete.

Reducing profanity. We remove subreddits that are not safe for work[3] and where 10%+ of comments include offensive/adult and vulgar tokens. We remove remaining comments that include offensive/adult tokens. Vulgar comments are preserved as we believe they are central to learning about negative emotions. The dataset includes the list of filtered tokens.

Manual review. We manually review identity comments and remove those offensive towards a particular ethnicity, gender, sexual orientation, or disability, to the best of our judgment.

Length filtering. We apply NLTK's word tokenizer and select comments 3-30 tokens long, including punctuation. To create a relatively balanced distribution of comment length, we perform downsampling, capping by the number of comments with the median token count (12).

Sentiment balancing. We reduce sentiment bias by removing subreddits with little representation of positive, negative, ambiguous, or neutral sentiment. To estimate a comment's sentiment, we run our emotion prediction model, trained on a pilot batch of 2.2k annotated examples. The mapping of emotions into sentiment categories is found in Figure 2. We exclude subreddits consisting of more than 30% neutral comments or less than 20% of negative, positive, or ambiguous comments.

Emotion balancing. We assign a predicted emotion to each comment using the pilot model described above. Then, we reduce emotion bias by downsampling the weakly-labelled data, capping by the number of comments belonging to the median emotion count.

Subreddit balancing. To avoid over representation of popular subreddits, we perform downsampling, capping by the median subreddit count.

From the remaining 315k comments (from 482 subreddits), we randomly sample for annotation.

Masking. We mask proper names referring to people with a [NAME] token, using a BERT-based Named Entity Tagger (Tsai et al., 2019). We mask religion terms with a [RELIGION] token. The list of these terms is included with our dataset. Note that raters viewed unmasked comments during rating.

3.2 Taxonomy of Emotions

When creating the taxonomy, we seek to jointly maximize the following objectives.

1. *Provide greatest coverage in terms of emotions expressed in our data.* To address this, we manually labeled a small subset of the data, and ran a pilot task where raters can suggest emotion labels on top of the pre-defined set.

[2] https://github.com/dewarim/reddit-data-tools

[3] http://redditlist.com/nsfw

2. *Provide greatest coverage in terms of kinds of emotional expression.* We consult psychology literature on emotion expression and recognition (Plutchik, 1980; Cowen and Keltner, 2017; Cowen et al., 2019b). Since, to our knowledge, there has not been research that identifies principal categories for emotion recognition in the domain of text (see Section 2.2), we consider those emotions that are identified as basic in other domains (video and speech) and that we can assume to apply to text as well.

3. *Limit overlap among emotions and limit the number of emotions.* We do not want to include emotions that are too similar, since that makes the annotation task more difficult. Moreover, combining similar labels with high coverage would result in an explosion in annotated labels.

The final set of selected emotions is listed in Table 4, and Figure 1. See Appendix B for more details on our multi-step taxonomy selection procedure.

3.3 Annotation

We assigned three raters to each example. For those examples where no raters agree on at least one emotion label, we assigned two additional raters. All raters are native English speakers from India.[4]

Instructions. Raters were asked to identify the emotions expressed by the writer of the text, given pre-defined emotion definitions (see Appendix A) and a few example texts for each emotion. Raters were free to select multiple emotions, but were asked to only select those ones for which they were reasonably confident that it is expressed in the text. If raters were not certain about any emotion being expressed, they were asked to select Neutral. We included a checkbox for raters to indicate if an example was particularly difficult to label, in which case they could select no emotions. We removed all examples for which no emotion was selected.

The rater interface. Reddit comments were presented with no additional metadata (such as the author or subreddit). To help raters navigate the large space of emotion in our taxonomy, they were presented a table containing all emotion categories aggregated by sentiment (by the mapping in Figure 2) and whether that emotion is generally expressed towards something (e.g. disapproval) or is more of an intrinsic feeling (e.g. joy). The instructions highlighted that this separation of categories was by no means clear-cut, but captured general tendencies, and we encouraged raters to ignore the categorization whenever they saw fit. Emotions with a straightforward mapping onto emojis were shown with an emoji in the UI, to further ease their interpretation.

[4]Cowen et al. (2019b) find that emotion judgments in Indian and US English speakers largely occupy the same dimensions.

Number of examples	58,009
Number of emotions	27 + neutral
Number of unique raters	82
Number of raters / example	3 or 5
Marked unclear or difficult to label	1.6%
Number of labels per example	1: 83% 2: 15% 3: 2% 4+: .2%
Number of examples w/ 2+ raters agreeing on at least 1 label	54,263 (94%)
Number of examples w/ 3+ raters agreeing on at least 1 label	17,763 (31%)

Table 2: Summary statistics of our labeled data.

4 Data Analysis

Table 2 shows summary statistics for the data. Most of the examples (83%) have a single emotion label and have at least two raters agreeing on a single label (94%). The Neutral category makes up 26% of all emotion labels – we exclude that category from the following analyses, since we do not consider it to be part of the semantic space of emotions.

Figure 1 shows the distribution of emotion labels. We can see a large disparity in terms of emotion frequencies (e.g. *admiration* is 30 times more frequent than *grief*), despite our emotion and sentiment balancing steps taken during data selection. This is expected given the disparate frequencies of emotions in natural human expression.

4.1 Interrater Correlation

We estimate rater agreement for each emotion via interrater correlation (Delgado and Tibau, 2019).[5] For each rater $r \in R$, we calculate the Spearman correlation between r's judgments and the mean

[5]We use correlations as opposed to Cohen's kappa (Cohen, 1960) because the former is a more interpretable metric and it is also more suitable for measuring agreement among a variable number of raters rating different examples. In Appendix C we report Cohen's kappa values as well, which correlate highly with the values obtained from interrater correlation (Pearson $r = 0.85$, $p < 0.001$).

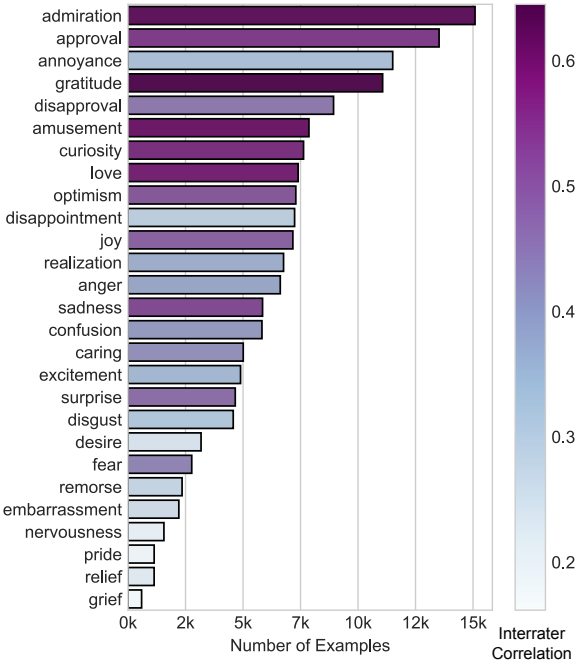

Figure 1: Our emotion categories, ordered by the number of examples where at least one rater uses a particular label. The color indicates the interrater correlation.

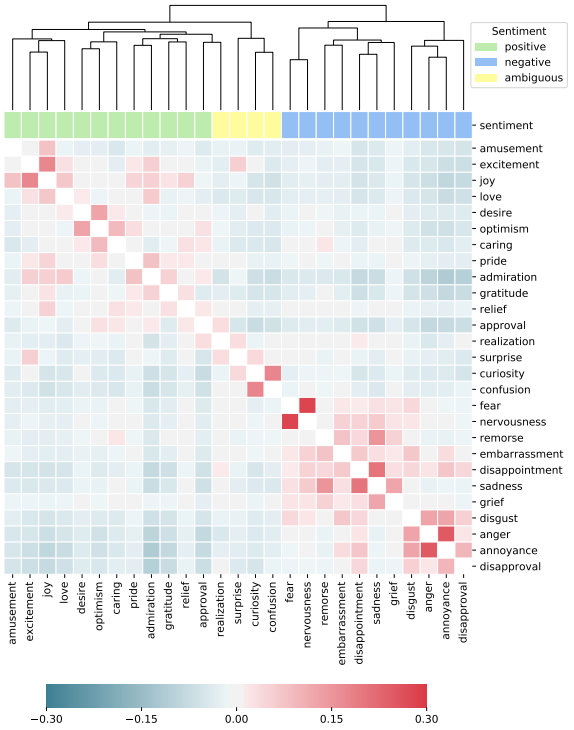

Figure 2: The heatmap shows the correlation between ratings for each emotion. The dendrogram represents the a hierarchical clustering of the ratings. The sentiment labeling was done *a priori* and it shows that the clusters closely map onto sentiment groups.

of other raters' judgments, for all examples that r rated. We then take the average of these rater-level correlation scores. In Section 4.3, we show that each emotion has significant interrater correlation, after controlling for several potential confounds.

Figure 1 shows that *gratitude*, *admiration* and *amusement* have the highest and *grief* and *nervousness* have the lowest interrater correlation. Emotion frequency correlates with interrater agreement but the two are not equivalent. Infrequent emotions can have relatively high interrater correlation (e.g., *fear*), and frequent emotions can have have relatively low interrater correlation (e.g., *annoyance*).

4.2 Correlation Among Emotions

To better understand the relationship between emotions in our data, we look at their correlations. Let N be the number of examples in our dataset. We obtain N dimensional vectors for each emotion by averaging raters' judgments for all examples labeled with that emotion. We calculate Pearson correlation values between each pair of emotions. The heatmap in Figure 2 shows that emotions that are related in intensity (e.g. *annoyance* and *anger*, *joy* and *excitement*, *nervousness* and *fear*) have a strong positive correlation. On the other hand, emotions that have the opposite sentiment are negatively correlated.

We also perform hierarchical clustering to uncover the nested structure of our taxonomy. We use correlation as a distance metric and ward as a linkage method, applied to the averaged ratings. The dendrogram on the top of Figure 2 shows that emotions that are related by intensity are neighbors, and that larger clusters map closely onto sentiment categories. Interestingly, emotions that we labeled as "ambiguous" in terms of sentiment (e.g. *surprise*) are closer to the positive than to the negative category. This suggests that in our data, ambiguous emotions are more likely to occur in the context of positive sentiment than that of negative sentiment.

4.3 Principal Preserved Component Analysis

To better understand agreement among raters and the latent structure of the emotion space, we apply Principal Preserved Component Analysis (PPCA) (Cowen et al., 2019b) to our data. PPCA extracts linear combinations of attributes (here, emotion judgments), that maximally covary across two sets of data that measure the same attributes (here, randomly split judgments for each example). Thus, PPCA allows us to uncover latent dimensions of

Algorithm 1 Leave-One-Rater-Out PPCA

$R \leftarrow$ set of raters
$E \leftarrow$ set of emotions
$C \in \mathbb{R}^{|R| \times |E|}$
for all raters $r \in \{1, ..., |R|\}$ **do**
  $n \leftarrow$ number of examples annotated by r
  $J \in \mathbb{R}^{n \times |R| \times |E|} \leftarrow$ all ratings for the examples annotated by r
  $J^{-r} \in \mathbb{R}^{n \times |R|-1 \times |E|} \leftarrow$ all ratings in J, excluding r
  $J^{r} \in \mathbb{R}^{n \times |E|} \leftarrow$ all ratings by r
  $X, Y \in \mathbb{R}^{n \times |E|} \leftarrow$ randomly split J^{-r} and average ratings across raters for both sets
  $W \in \mathbb{R}^{|E| \times |E|} \leftarrow$ result of $PPCA(X, Y)$
  for all components† $\mathbf{w}_{i \in \{1,...,|E|\}}$ in W **do**
    $\mathbf{v}_i^r \leftarrow$ projection‡ of J^r onto $\mathbf{w}_i$
    $\mathbf{v}_i^{-r} \leftarrow$ projection‡ of J^{-r} onto $\mathbf{w}_i$
    $C_{r,i} \leftarrow$ correlation between $\mathbf{v}_i^r$ and $\mathbf{v}_i^{-r}$, partialing out $\mathbf{v}_k^{-r} \forall k \in \{1, ..., i-1\}$
  end for
end for
$C' \leftarrow$ Wilcoxon signed rank test on C
$C'' \leftarrow$ Bonferroni correction on $C'(\alpha = 0.05)$

†in descending order of eigenvalue
‡we demean vectors before projection

emotion that have high agreement across raters.

Unlike Principal Component Analysis (PCA), PPCA examines the cross-covariance between datasets rather than the variance–covariance matrix within a single dataset. We obtain the principal preserved components (PPCs) of two datasets (matrices) $X, Y \in \mathbb{R}^{N \times |E|}$, where N is the number of examples and $|E|$ is the number of emotions, by calculating the *eigenvectors* of the symmetrized cross covariance matrix $X^T Y + Y^T X$.

Extracting significant dimensions. We remove examples labeled as Neutral, and keep those examples that still have at least 3 ratings after this filtering step. We then determine the number of significant dimensions using a leave-one-rater out analysis, as described by Algorithm 1.

We find that all 27 PPCs are highly significant. Specifically, Bonferroni-corrected p-values are less than 1.5e-6 for all dimensions (corrected α = 0.0017), suggesting that the emotions were highly dissociable. Such a high degree of significance for all dimensions is nontrivial. For example, Cowen et al. (2019b) find that only 12 out of their 30 emotion categories are significantly dissociable.

t-SNE projection. To better understand how the examples are organized in the emotion space, we apply t-SNE, a dimension reduction method that seeks to preserve distances between data points, using the scikit-learn package (Pedregosa et al., 2011). The dataset can be explored in our interactive plot[6], where one can also look at the texts and the annotations. The color of each data point is the weighted average of the RGB values representing those emotions that at least half of the raters selected.

4.4 Linguistic Correlates of Emotions

We extract the lexical correlates of each emotion by calculating the log odds ratio, informative Dirichlet prior (Monroe et al., 2008) of all tokens for each emotion category contrasting to all other emotions. Since the log odds are z-scored, all values greater than 3 indicate highly significant (>3 std) association with the corresponding emotion. We list the top 5 tokens for each category in Table 3. We find that those emotions that are highly significantly associated with certain tokens (e.g. *gratitude* with "thanks", *amusement* with "lol") tend to have the highest interrater correlation (see Figure 1). Conversely, emotions that have fewer significantly associated tokens (e.g. *grief* and *nervousness*) tend to have low interrater correlation. These results suggest certain emotions are more verbally implicit and may require more context to be interpreted.

5 Modeling

We present a strong baseline emotion prediction model for GoEmotions.

5.1 Data Preparation

To minimize the noise in our data, we filter out emotion labels selected by only a single annotator. We keep examples with at least one label after this filtering is performed — this amounts to 93% of the original data. We randomly split this data into train (80%), dev (10%) and test (10%) sets. We only evaluate on the test set once the model is finalized.

Even though we filter our data for the baseline experiments, we see particular value in the 4K examples that lack agreement. This subset of the data likely contains edge/difficult examples for the emotion domain (e.g., emotion-ambiguous text), and present challenges for further exploration. That is

[6] https://nlp.stanford.edu/~ddemszky/goemotions/tsne.html

admiration	amusement	approval	caring	anger	annoyance	disappointment	disapproval	confusion
great (42)	lol (66)	agree (24)	you (12)	fuck (24)	annoying (14)	disappointing (11)	not (16)	confused (18)
awesome (32)	haha (32)	not (13)	worry (11)	hate (18)	stupid (13)	disappointed (10)	don't (14)	why (11)
amazing (30)	funny (27)	don't (12)	careful (9)	fucking (18)	fucking (12)	bad (9)	disagree (9)	sure (10)
good (28)	lmao (21)	yes (12)	stay (9)	angry (11)	shit (10)	disappointment (7)	nope (8)	what (10)
beautiful (23)	hilarious (18)	agreed (11)	your (8)	dare (10)	dumb (9)	unfortunately (7)	doesn't (7)	understand (8)
desire	**excitement**	**gratitude**	**joy**	**disgust**	**embarrassment**	**fear**	**grief**	**curiosity**
wish (29)	excited (21)	thanks (75)	happy (32)	disgusting (22)	embarrassing (12)	scared (16)	died (6)	curious (22)
want (8)	happy (8)	thank (69)	glad (27)	awful (14)	shame (11)	afraid (16)	rip (4)	what (18)
wanted (6)	cake (8)	for (24)	enjoy (20)	worst (13)	awkward (10)	scary (15)		why (13)
could (6)	wow (8)	you (18)	enjoyed (12)	worse (12)	embarrassment (8)	terrible (12)		how (11)
ambitious (4)	interesting (7)	sharing (17)	fun (12)	weird (9)	embarrassed (7)	terrifying (11)		did (10)
love	**optimism**	**pride**	**relief**	**nervousness**	**remorse**	**sadness**	**realization**	**surprise**
love (76)	hope (45)	proud (14)	glad (5)	nervous (8)	sorry (39)	sad (31)	realize (14)	wow (23)
loved (21)	hopefully (19)	pride (4)	relieved (4)	worried (8)	regret (9)	sadly (16)	realized (12)	surprised (21)
favorite (13)	luck (18)	accomplishment	relieving (4)	anxiety (6)	apologies (7)	sorry (15)	realised (7)	wonder (15)
loves (12)	hoping (16)	(4)	relief (4)	anxious (4)	apologize (6)	painful (10)	realization (6)	shocked (12)
like (9)	will (8)			worrying (4)	guilt (5)	crying (9)	thought (6)	omg (11)

Table 3: Top 5 words associated with each emotion (positive , negative , ambiguous). The rounded z-scored log odds ratios in the parentheses, with the threshold set at 3, indicate significance of association.

why we release all 58K examples with all annotators' ratings.

Grouping emotions. We create a hierarchical grouping of our taxonomy, and evaluate the model performance on each level of the hierarchy. A sentiment level divides the labels into 4 categories – *positive*, *negative*, *ambiguous* and Neutral – with the Neutral category intact, and the rest of the mapping as shown in Figure 2. The Ekman level further divides the taxonomy using the Neutral label and the following 6 groups: *anger* (maps to: *anger*, *annoyance*, *disapproval*), *disgust* (maps to: disgust), *fear* (maps to: *fear*, *nervousness*), *joy* (all *positive* emotions), *sadness* (maps to: *sadness*, *disappointment*, *embarrassment*, *grief*, *remorse*) and *surprise* (all *ambiguous* emotions).

5.2 Model Architecture

We use the BERT-base model (Devlin et al., 2019) for our experiments. 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. As an additional baseline, we train a bidirectional LSTM.

5.3 Parameter Settings

When finetuning the pre-trained BERT model, we keep most of the hyperparameters set by Devlin et al. (2019) intact and only change the batch size and learning rate. We find that training for at least 4 epochs is necessary for learning the data, but training for more epochs results in overfitting. We also find that a small batch size of 16 and learning rate of 5e-5 yields the best performance.

For the biLSTM, we set the hidden layer dimensionality to 256, the learning rate to 0.1, with a decay rate of 0.95. We apply a dropout of 0.7.

5.4 Results

Table 4 summarizes the performance of our best model, BERT, on the test set, which achieves an average F1-score of .46 (std=.19). The model obtains the best performance on emotions with overt lexical markers, such as *gratitude* (.86), *amusement* (.8) and *love* (.78). The model obtains the lowest F1-score on *grief* (0), *relief* (.15) and *realization* (.21), which are the lowest frequency emotions. We find that less frequent emotions tend to be confused by the model with more frequent emotions related in sentiment and intensity (e.g., *grief* with *sadness*, *pride* with *admiration*, *nervousness* with *fear*) — see Appendix G for a more detailed analysis.

Table 5 and Table 6 show results for a sentiment-grouped model (F1-score = .69) and an Ekman-grouped model (F1-score = .64), respectively. The significant performance increase in the transition from full to Ekman-level taxonomy indicates that this grouping mitigates confusion among inner-group lower-level categories.

The biLSTM model performs significantly worse than BERT, obtaining an average F1-score of .41 for the full taxonomy, .53 for an Ekman-grouped model and .6 for a sentiment-grouped model.

6 Transfer Learning Experiments

We conduct transfer learning experiments on existing emotion benchmarks, in order to show our data generalizes across domains and taxonomies. The goal is to demonstrate that given little labeled data in a target domain, one can utilize GoEmotions as baseline emotion understanding data.

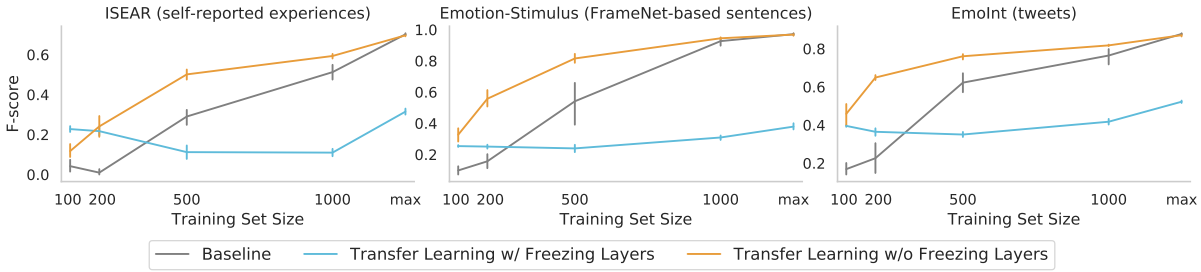

Figure 3: Transfer learning results in terms of average F1-scores across emotion categories. The bars indicate the 95% confidence intervals, which we obtain from 10 different runs on 10 different random splits of the data.

Emotion	Precision	Recall	F1
admiration	0.53	0.83	0.65
amusement	0.70	0.94	0.80
anger	0.36	0.66	0.47
annoyance	0.24	0.63	0.34
approval	0.26	0.57	0.36
caring	0.30	0.56	0.39
confusion	0.24	0.76	0.37
curiosity	0.40	0.84	0.54
desire	0.43	0.59	0.49
disappointment	0.19	0.52	0.28
disapproval	0.29	0.61	0.39
disgust	0.34	0.66	0.45
embarrassment	0.39	0.49	0.43
excitement	0.26	0.52	0.34
fear	0.46	0.85	0.60
gratitude	0.79	0.95	0.86
grief	0.00	0.00	0.00
joy	0.39	0.73	0.51
love	0.68	0.92	0.78
nervousness	0.28	0.48	0.35
neutral	0.56	0.84	0.68
optimism	0.41	0.69	0.51
pride	0.67	0.25	0.36
realization	0.16	0.29	0.21
relief	0.50	0.09	0.15
remorse	0.53	0.88	0.66
sadness	0.38	0.71	0.49
surprise	0.40	0.66	0.50
macro-average	**0.40**	**0.63**	**0.46**
std	**0.18**	**0.24**	**0.19**

Table 4: Results based on GoEmotions taxonomy.

6.1 Emotion Benchmark Datasets

We consider the nine benchmark datasets from Bostan and Klinger (2018)'s Unified Dataset, which vary in terms of their size, domain, quality and taxonomy. In the interest of space, we only discuss three of these datasets here, chosen based on their diversity of domains. In our experiments, we observe similar trends for the additional benchmarks, and all are included in the Appendix H.

Sentiment	Precision	Recall	F1
ambiguous	0.54	0.66	0.60
negative	0.65	0.76	0.70
neutral	0.64	0.69	0.67
positive	0.78	0.87	0.82
macro-average	**0.65**	**0.74**	**0.69**
std	**0.09**	**0.10**	**0.09**

Table 5: Results based on sentiment-grouped data.

Ekman Emotion	Precision	Recall	F1
anger	0.50	0.65	0.57
disgust	0.52	0.53	0.53
fear	0.61	0.76	0.68
joy	0.77	0.88	0.82
neutral	0.66	0.67	0.66
sadness	0.56	0.62	0.59
surprise	0.53	0.70	0.61
macro-average	**0.59**	**0.69**	**0.64**
std	**0.10**	**0.11**	**0.10**

Table 6: Results using Ekman's taxonomy.

The International Survey on Emotion Antecedents and Reactions (ISEAR) (Scherer and Wallbott, 1994) is a collection of personal reports on emotional events, written by 3000 people from different cultural backgrounds. The dataset contains 8k sentences, each labeled with a single emotion. The categories are *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*.

EmoInt (Mohammad et al., 2018) is part of the SemEval 2018 benchmark, and it contains crowdsourced annotations for 7k tweets. The labels are

intensity annotations for *anger*, *joy*, *sadness*, and *fear*. We obtain binary annotations for these emotions by using .5 as the cutoff.

Emotion-Stimulus (Ghazi et al., 2015) contains annotations for 2.4k sentences generated based on FrameNet's emotion-directed frames. Their taxonomy is *anger*, *disgust*, *fear*, *joy*, *sadness*, *shame* and *surprise*.

6.2 Experimental Setup

Training set size. We experiment with varying amount of training data from the target domain dataset, including 100, 200, 500, 1000, and 80% (named "max") of dataset examples. We generate 10 random splits for each train set size, with the remaining examples held as a test set.

We report the results of the finetuning experiments detailed below for each data size, with confidence intervals based on repeated experiments using the splits.

Finetuning. We compare three different finetuning setups. In the BASELINE setup, we finetune BERT only on the target dataset. In the FREEZE setup, we first finetune BERT on GoEmotions, then perform transfer learning by replacing the final dense layer, freezing all layers besides the last layer and finetuning on the target dataset. The NOFREEZE setup is the same as FREEZE, except that we do not freeze the bottom layers. We hold the batch size at 16, learning rate at 2e-5 and number of epochs at 3 for all experiments.

6.3 Results

The results in Figure 3 suggest that our dataset generalizes well to different domains and taxonomies, and that using a model using GoEmotions can help in cases when there is limited data from the target domain, or limited resources for labeling.

Given limited target domain data (100 or 200 examples), both FREEZE and NOFREEZE yield significantly higher performance than the BASELINE, for all three datasets. Importantly, NOFREEZE results show significantly higher performance for all training set sizes, except for "max", where NOFREEZE and BASELINE perform similarly.

7 Conclusion

We present GoEmotions, a large, manually annotated, carefully curated dataset for fine-grained emotion prediction. We provide a detailed data analysis, demonstrating the reliability of the annotations for the full taxonomy. We show the generalizability of the data across domains and taxonomies via transfer learning experiments. We build a strong baseline by fine-tuning a BERT model, however, the results suggest much room for future improvement. Future work can explore the cross-cultural robustness of emotion ratings, and extend the taxonomy to other languages and domains.

Data Disclaimer: We are aware that the dataset contains biases and is not representative of global diversity. We are aware that the dataset contains potentially problematic content. Potential biases in the data include: Inherent biases in Reddit and user base biases, the offensive/vulgar word lists used for data filtering, inherent or unconscious bias in assessment of offensive identity labels, annotators were all native English speakers from India. All these likely affect labeling, precision, and recall for a trained model. The emotion pilot model used for sentiment labeling, was trained on examples reviewed by the research team. Anyone using this dataset should be aware of these limitations of the dataset.

Acknowledgments

We thank the three anonymous reviewers for their constructive feedback. We would also like to thank the annotators for their hard work.

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.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the

wild: A case for locating elusive phenomena in social media posts. 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 1664–1674.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. 2019a. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90.

Alan S Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. 2018. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698–712.

Alan S Cowen, Xia Fang, Disa Sauter, and Dacher Keltner. in press. What music makes us feel: At least thirteen dimensions organize subjective experiences associated with music across cultures. *Proceedings of the National Academy of Sciences*.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Alan S Cowen and Dacher Keltner. 2019. What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*.

Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019b. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, 3(4):369.

CrowdFlower. 2016. https://www.figure-eight.com/data/sentiment-analysis-emotion-text/.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Maeve Duggan and Aaron Smith. 2013. 6% of online adults are reddit users. *Pew Internet & American Life Project*, 3:1–10.

Paul Ekman. 1992a. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Paul Ekman. 1992b. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting Emotion Stimuli in Emotion-Bearing Sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Chao-Chun Hsu and Lun-Wei Ku. 2018. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 27–31, Melbourne, Australia. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. Dens: A dataset for multi-class emotion analysis. *arXiv preprint arXiv:1910.11769*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M Mohammad. 2012. # emotional tweets. In *Proceedings of 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*, pages 246–255. Association for Computational Linguistics.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Rosalind W Picard. 1997. *Affective Computing*. MIT Press.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of personality and social psychology*, 66(2):310.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Association for Computational Linguistics*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. In *EMNLP 2019*.

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. IEEE.

Cebrian-M. Yanardag, P. and I. Rahwan. 2018. Norman: World's first psychopath ai.

A Emotion Definitions

admiration 👏 Finding something impressive or worthy of respect.
amusement 😂 Finding something funny or being entertained.
anger 😡 A strong feeling of displeasure or antagonism.
annoyance 😒 Mild anger, irritation.
approval 👍 Having or expressing a favorable opinion.
caring Displaying kindness and concern for others.
confusion 😕 Lack of understanding, uncertainty.
curiosity A strong desire to know or learn something.
desire A strong feeling of wanting something or wishing for something to happen.
disappointment Sadness or displeasure caused by the nonfulfillment of one's hopes or expectations.
disapproval 👎 Having or expressing an unfavorable opinion.
disgust 🤮 Revulsion or strong disapproval aroused by something unpleasant or offensive.
embarrassment 😳 Self-consciousness, shame, or awkwardness.
excitement 🤩 Feeling of great enthusiasm and eagerness.
fear 😨 Being afraid or worried.
gratitude 🙏 A feeling of thankfulness and appreciation.
grief Intense sorrow, especially caused by someone's death.
joy 😃 A feeling of pleasure and happiness.
love ❤️ A strong positive emotion of regard and affection.
nervousness Apprehension, worry, anxiety.
optimism 🤞 Hopefulness and confidence about the future or the success of something.
pride Pleasure or satisfaction due to ones own achievements or the achievements of those with whom one is closely associated.
realization Becoming aware of something.
relief Reassurance and relaxation following release from anxiety or distress.
remorse Regret or guilty feeling.
sadness 😢 Emotional pain, sorrow.
surprise 😮 Feeling astonished, startled by something unexpected.

B Taxonomy Selection & Data Collection

We selected our taxonomy through a careful multi-round process. In the first pilot round of data collection, we used emotions that were identified to be salient by Cowen and Keltner (2017), making sure that our set includes Ekman's emotion categories, as used in previous NLP work. In this round, we also included an open input box where annotators could suggest emotion(s) that were not among the options. We annotated 3K examples in the first round. We updated the taxonomy based on the results of this round (see details below). In the second pilot round of data collection, we repeated this process with 2k new examples, once again updating the taxonomy.

While reviewing the results from the pilot rounds, we identified and removed emotions that were scarcely selected by annotators and/or had low interrater agreement due to being very similar to other emotions or being too difficult to detect from text. These emotions were *boredom*, *doubt*, *heartbroken*, *indifference* and *calmness*. We also identified and added those emotions to our taxonomy that were frequently suggested by raters and/or seemed to be represented in the data upon manual inspection. These emotions were *desire*, *disappointment*, *pride*, *realization*, *relief* and *remorse*. In this process, we also refined the category names (e.g. replacing *ecstasy* with *excitement*), to ones that seemed interpretable to annotators. This is how we arrived at the final set of 27 emotions + Neutral. Our high interrater agreement in the final data can be partially explained by the fact that we took interpretability into consideration while constructing the taxonomy. The dataset is we are releasing was labeled in the third round over the final taxonomy.

C Cohen's Kappa Values

In Section 4.1, we measure agreement between raters via Spearman correlation, following considerations by Delgado and Tibau (2019). In Table 7, we report the Cohen's kappa values for comparison, which we obtain by randomly sampling two ratings for each example and calculating the Cohen's kappa between these two sets of ratings. We find that all Cohen's kappa values are greater than 0, showing rater agreement. Moreover, the Cohen's kappa values correlate highly with the interrater correlation values (Pearson $r = 0.85$, $p < 0.001$), providing corroborative evidence for the significant degree of interrater agreement for each emotion.

Emotion	Interrater Correlation	Cohen's kappa
admiration	0.535	0.468
amusement	0.482	0.474
anger	0.207	0.307
annoyance	0.193	0.192
approval	0.385	0.187
caring	0.237	0.252
confusion	0.217	0.270
curiosity	0.418	0.366
desire	0.177	0.251
disappointment	0.186	0.184
disapproval	0.274	0.234
disgust	0.192	0.241
embarrassment	0.177	0.218
excitement	0.193	0.222
fear	0.266	0.394
gratitude	0.645	0.749
grief	0.162	0.095
joy	0.296	0.301
love	0.446	0.555
nervousness	0.164	0.144
optimism	0.322	0.300
pride	0.163	0.148
realization	0.194	0.155
relief	0.172	0.185
remorse	0.178	0.358
sadness	0.346	0.336
surprise	0.275	0.331

Table 7: Interrater agreement, as measured by interrater correlation and Cohen's kappa

D Sentiment of Reddit Subreddits

In Section 3, we describe how we obtain subreddits that are balanced in terms of sentiment. Here, we note the distribution of sentiments across subreddits *before* we apply the filtering: neutral (M=28%, STD=11%), positive (M=41%, STD=11%), negative (M=19%, STD=7%), ambiguous (M=35%, STD=8%). After filtering, the distribution of sentiments across our remaining subreddits became: neutral (M=24%, STD=5%), positive (M=35%, STD=6%), negative (M=27%, STD=4%), ambiguous (M=33%, STD=4%).

E BERT's Most Activated Layers

To better understand whether there are any layers in BERT that are particularly important for our task, we freeze BERT and calculate the center of gravity (Tenney et al., 2019) based on scalar mixing weights (Peters et al., 2018). We find that all layers are similarly important for our task, with center of gravity = 6.19 (see Figure 4). This is consistent with Tenney et al. (2019), who have also found that tasks involving high-level semantics tend to make use of all BERT layers.

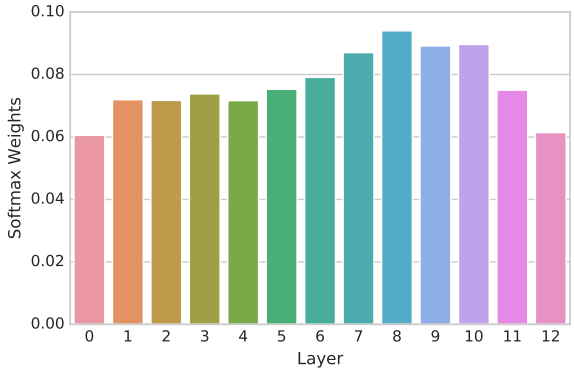

Figure 4: Softmax weights of each BERT layer when trained on our dataset.

F Number of Emotion Labels Per Example

Figure 5 shows the number of emotion labels per example before and after we filter for those labels that have agreement. We use the filtered set of labels for training and testing our models.

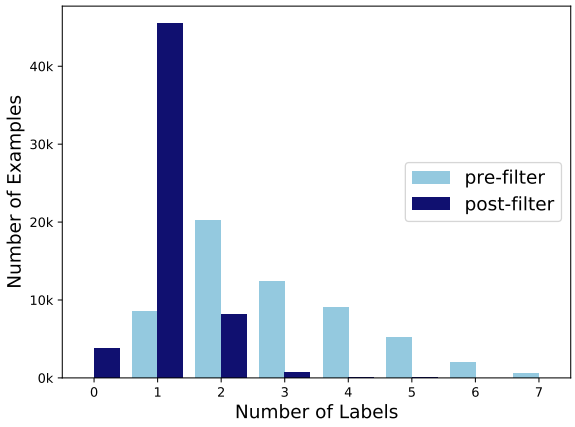

Figure 5: Number of emotion labels per example before and after filtering the labels chosen by only a single annotator.

G Confusion Matrix

Figure 6 shows the normalized confusion matrix for our model predictions. Since GoEmotions is a multilabel dataset, we calculate the confusion matrix

similarly as we would calculate a co-occurrence matrix: for each true label, we increase the count for each predicted label. Specifically, we define a matrix M where $M_{i,j}$ denotes the raw confusion count between the true label i and the predicted label j. For example, if the true labels are *joy* and *admiration*, and the predicted labels are *joy* and *pride*, then we increase the count for $M_{joy,joy}$, $M_{joy,pride}$, $M_{admiration,joy}$ and $M_{admiration,pride}$. In practice, since most of our examples only has a single label (see Figure 5), our confusion matrix is very similar to one calculated for a single-label classification task.

Given the disparate frequencies among the labels, we normalize M by dividing the counts in each row (representing counts for each true emotion label) by the sum of that row. The heatmap in Figure 6 shows these normalized counts. We find that the model tends to confuse emotions that are related in sentiment and intensity (e.g., *grief* and *sadness*, *pride* and *admiration*, *nervousness* and *fear*).

We also perform hierarchical clustering over the normalized confusion matrix using correlation as a distance metric and ward as a linkage method. We find that the model learns relatively similar clusters as the ones in Figure 2, even though the training data only includes a subset of the labels that have agreement (see Figure 5).

H Transfer Learning Results

Figure 7 shows the results for all 9 datasets that are downloadable and have categorical emotions in the Unified Dataset (Bostan and Klinger, 2018). These datasets are DailyDialog (Li et al., 2017), Emotion-Stimulus (Ghazi et al., 2015), Affective Text (Strapparava and Mihalcea, 2007), CrowdFlower (CrowdFlower, 2016), Electoral Tweets (Mohammad et al., 2015), ISEAR (Scherer and Wallbott, 1994), the Twitter Emotion Corpus (TEC) (Mohammad, 2012), EmoInt (Mohammad et al., 2018) and the Stance Sentiment Emotion Corpus (SSEC) (Schuff et al., 2017).

We describe the experimental setup in Section 6.2, which we use across all datasets. We find that transfer learning helps in the case of all datasets, especially when there is limited training data. Interestingly, in the case of CrowdFlower, which is known to be noisy (Bostan and Klinger, 2018) and Electoral Tweets, which is a small dataset of ∼4k labeled examples and a large taxonomy of 36 emotions, FREEZE gives a significant boost of performance over the BASELINE and NOFREEZE for all training set sizes besides "max".

For the other datasets, we find that FREEZE tends to give a performance boost compared to the other setups only up to a couple of hundred training examples. For 500-1000 training examples, NOFREEZE tends to outperform the BASELINE, but we can see that these two setups come closer when there is more training data available. These results suggests that our dataset helps if there is limited data from the target domain.

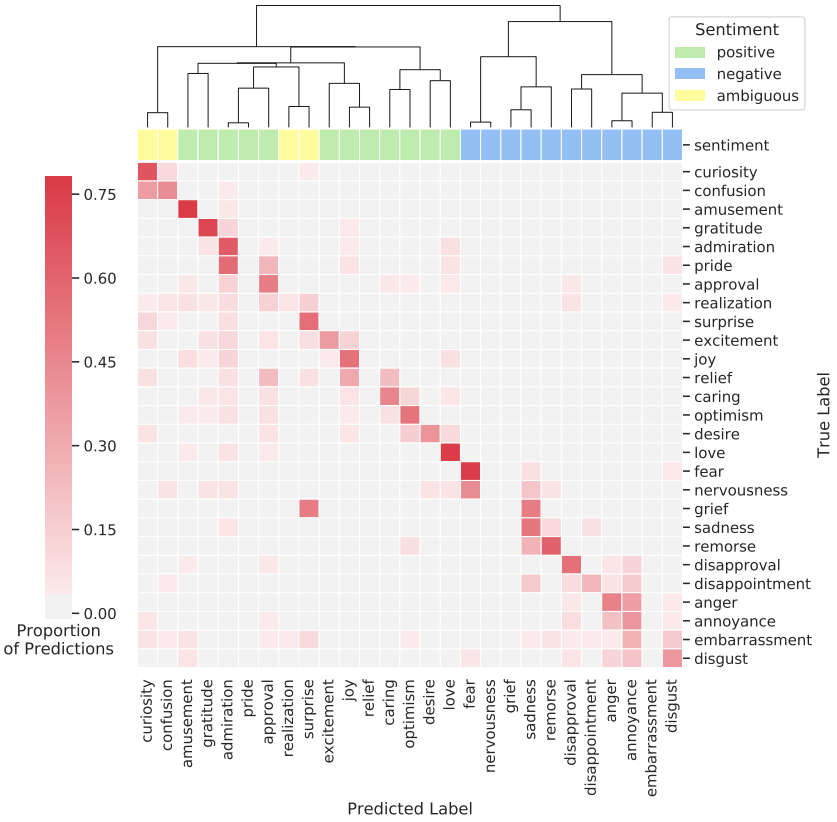

Figure 6: A normalized confusion matrix for our model predictions. The plot shows that the model confuses emotions with other emotions that are related in intensity and sentiment.

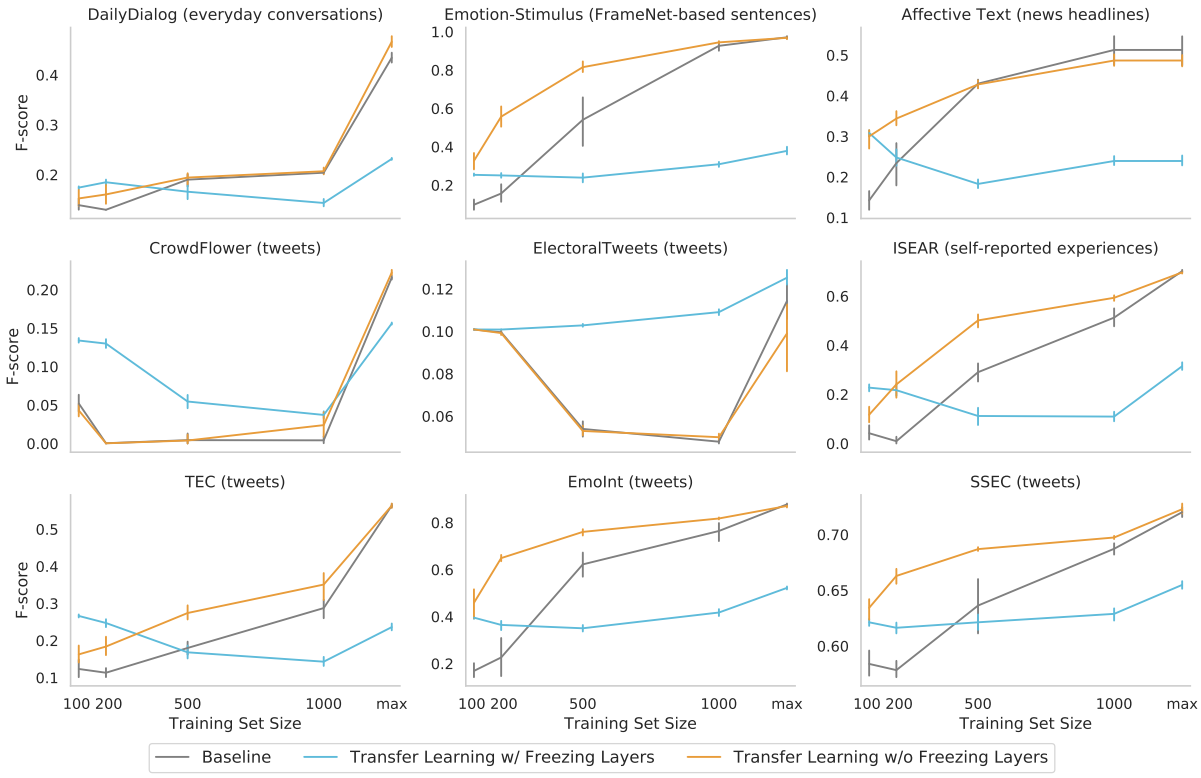

Figure 7: Transfer learning results on 9 emotion benchmarks from the Unified Dataset (Bostan and Klinger, 2018).