GoEmotions Model Card

Author: Dana Movshovitz-Attias
Latest update: 4 Dec, 2020

Based on Model Cards for Model Reporting, In Proceedings of FAT Conference (FAT*2019). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3287560.3287596*

1. Qualitative Information

1a. Model Details

This card describes models built using the GoEmotions public dataset. **It does not describe a specific pre-trained model, but rather a model that will result from using the GoEmotions data.**

- Developed at Google Research.
- **Contact**: Dora Demszky, Dana Movshovitz-Attias
- **Data collection date**: 05-2019
- **Paper**: GoEmotions: A Dataset of Fine-Grained Emotions.
- **Model Type**: We describe here the BERT model trained over the data, as part of the accompanying paper. Whenever possible, this card will describe a general GoEmotions-based model.
- **Modeling, training, parameters**: Described in Section 5 of the paper.
- **Fairness constraints**: The collected data originated from Reddit 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 took a series of data curation measures to ensure our data does not reinforce general, nor emotion-specific, language biases. We list here a sample of the data curation measures taken. Full details can be found in Section 3 of the paper.

  - Reducing profanity: We remove subreddits that are not-safe-for-work and where 10%+ of comments include offensive/adult and vulgar tokens.
  - Sentiment/Emotion/Subreddit balancing: We subsample the data to adjust the representation of estimated sentiment/emotion categories and of specific subreddits.
  - NER masking: We mask people and religion terms using an NER tagging model.

- **License**: The data is released under the Apache 2.0 license, with the following disclaimers:
  - 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 labelling, 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.
- **Citation**:

```
@inproceedings{demszky-2020-goemotions,
       title = "{G}o{E}motions: A Dataset of Fine-Grained Emotions",
    author = "Demszky, Dorottya  and
      Movshovitz-Attias, Dana  and
      Ko, Jeongwoo  and
      Cowen, Alan  and
      Nemade, Gaurav  and
      Ravi, Sujith",
    booktitle = "Proceedings of the 58th Annual Meeting of the
Association for Computational Linguistics",
    month = jul,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://www.aclweb.org/anthology/2020.acl-main.372",
    pages = "4040--4054",
```

```
}
```

1b. Considerations When Using the Model

- Below are canonical uses for the GoEmotion data and any model based on it.

1.	**User Feedback**	- Understanding granular level emotion within short feedback given by users through forms, surveys, or other means.
2.	**Social Listening**	- Understanding granular level emotion of what people are thinking about products or services. - This can be done by evaluating isolated messages from Twitter, Reddit, YouTube comments and more.
3.	**Customer Support Interaction**	- Evaluating customer support data (call logs, transcripts, etc.) in order to understand customer emotions. - Identify topics and emotions associated with actionable responses.
4.	**Expressive Content**	- Understanding the granular level emotion of a user in a local point in the course of a conversation. - The can be used in order to empathetically react to that emotion.

- This model is particularly useful for:
    - Researchers: The data collected in GoEmotions is mainly targeted at advancing state-of-the-art modeling of emotion from text.
    - Developers: App developers may find it useful to recognize the localized emotion expressed by their users, thus enabling an affective product that can react empathetically to their users.
- Limitations:
    - The data is labeled at a localized level, with labels appearing on single input messages, without additional conversational context. The data was similarly labeled without additional context available to raters. For this reason:
    - The data can only be used to determine the local emotional state of a user.
    - Aggregating local emotion predictions does not guarantee a meaningful interpretation over a stream of data with regards to emotional state or mental well being.
    - Context-less emotion prediction means that the input text can often be interpreted in multiple fashions. As such, we expect the prediction to be ambiguous for many input texts (i.e., present multiple, possibly contradicting, predicted emotions). The interpretation of such ambiguous predictions means that each of the emotions predicted with high confidence are possibly expressed in the input text.

    - In the course of creating the GoEmotions data, we took great care to address potential biases presented in it, see section 3 in GoEmotions: A Dataset of Fine-Grained Emotions. Despite this, models trained over the data should still take active steps to address learned biases in the resulting models, through model de-biasing techniques.

2. Quantitative Analyses

2a. Factors

In the paper we present a BERT-base model trained over the GoEmotions data, and evaluated on similarly treated test data. We additionally test the ability of our data to predict emotions labeled in other emotion-labeled resources, in order to demonstrate the generalizability of the data. The following sections describe the model performance in all these environments.

- Groups:
    - GoEmotions eval set: Reddit users (data filtered and treated in a similar manner to training set).
    - EmoInt (Mohammad et al., 2018): Messages of Twitter users.
    - ISEAR (Scherer and Wallbott, 1994): A collection of personal reports on emotional events, written by 3000 people from different cultural backgrounds.
    - Emotion-Stimulus (Ghazi et al., 2015): contains annotations for 2.4k sentences generated based on FrameNet's emotion-directed frames.
- Platforms:
    - Social media: GoEmotions, EmoInt
    - Personal Recounts: ISEAR, Emotion-Stimulus

2b. Metrics

- Model performance:
    - Emotion-level Precision, Recall, F1: Measured per each emotion in the GoEmotions taxonomy.
    - Ekman-level Precision, Recall, F1: Aggregated at the level of Ekman taxonomy, which is commonly used in emotion literature.
    - Sentiment-level Precision, Recall, F1: Aggregated at the Sentiment level (positive, negative, neutral/ambiguous), which is commonly used in emotion literature.
    - Transfer learning: F1 score on data transferred between domain X and GoEmotions.
- Decision thresholds: No thresholds are used. The data is presented in full granularity.
- Uncertainty and variability: Repeated experiments have yielded results with similar taxonomical rankings.

2c. Evaluation Data

We evaluate our model on 10 publicly available datasets, including the GoEmotions eval set, and 9 benchmark datasets presented in the compilation work by Bostan and Klinger.

2e. Results

Below are the main tables and figures describing the evaluations results. Full details can be found in the paper.

We highlight here the differences in performance between emotion categories, as seen in Table 4. This can be attributed to the following:

- Data variations: Naturally, not all emotions are expressed uniformly. The vast variation in data availability of data per emotion also affects the model's ability to learn that emotion.
- Data processing: Our data curation process (described in section 1a) was designed to address potential biases in the raw model input. As part of this process, we may remain with some skew that affects the ability to learn extreme emotions, for example, due to removing a large amount of input including profanities.
- Linguistic correlates: Some emotions are more strongly associated with clear linguistic correlates, for example Gratitude and "thank you". These clear signals make some emotions easier to label than others, and in turn can affect classification results. We have a detailed analysis on this phenomena in the paper.

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.

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.

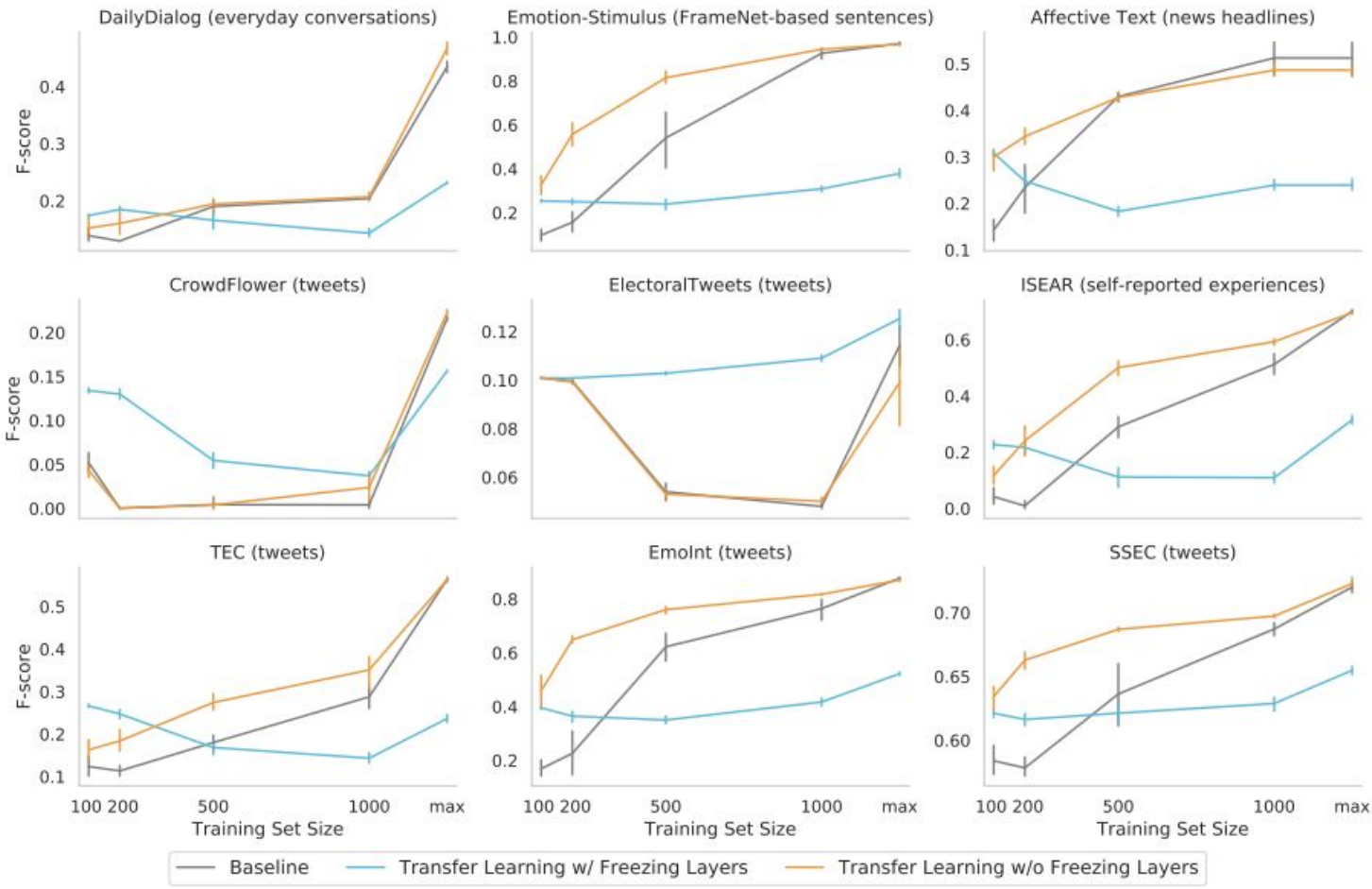

Figure 7: Transfer learning results on 9 emotion benchmarks from the Unified Dataset (Bostan and Klinger, 2018).

3. Considerations

3a. Ethical Considerations

- Data: The model and dataset are based on Reddit comments. Despite the authors' attempts to de-bias and remove sensitive aspects of the data, we are aware that it can potentially contain problematic content. See more License limitations in section 1a.
- Human life: This model is designed for localized prediction on emotion from input text. It is not designed for, nor suitable for, higher level understanding of mental health.
- Mitigations: During the development of the dataset we took a series of data curation measures to ensure it does not reinforce general, nor emotion-specific, language biases. Full details can be found in Section 3 of the paper. Any model build on top of this data should take specific care to address potential data-introduced biases.

3b. Examples of Use-case based Considerations

In this section we describe modeling considerations, with a sample use-case in mind. This is meant to tie together several points made in the text above.

Consider using an emotion-aware model for Social Listening (see section 1b for use-case description). The model developer (or a researcher) would have to take into account the data limitations. As an example, the predictions made by the emotion model are localized, and therefore are only reliable in the context of a single text input. This means that it is not recommended to use this data, or a model based on the data, to generate an explicit user emotion-profile with respect to a product or service. Similarly, the model's predictions are context-free, meaning that based solely on the data presented here, there may be missing factors for fully evaluating the *source* of an expressed emotion. To address this, the developer would have to combine additional relevant information, for example, recent world events, user geographical location, etc', that may affect the results.

In section 2e we highlight the difference in prediction quality with respect to different emotion categories, and reasons and insights around those differences. There are likely additional factors playing into the ability to correctly predict specific emotion categories, for example, those relating to the user demographics and types of expression exhibited by population subsets. For Social Listening, this can be manifested as unique linguistic expressions of positive/negative emotions towards a product, which vary by, say, demographics. Such unique expressions may not be well represented by our data, as it was collected from a set of users that are not necessarily world representative. In order to fully account for these, it is the responsibility of a developer using the data to incorporate these signals in order to ensure fair and unbiased predictions towards such demographics, or in order to inform the model of missing information for making reliable and confident predictions, in this case.