Label-Aware Attention for Fine-Grained Emotion Detection

Yilin Wang, Ethan Wu, Ziqi Liu
{yilinwan, yongyiw, ziqil2}@andrew.cmu.edu
Carnegie Mellon University

1.5em

Label-Aware Attention for Fine-Grained Emotion Detection

Yilin Wang, Ethan Wu, Ziqi Liu
{yilinwan, yongyiw, ziqil2}@andrew.cmu.edu
Carnegie Mellon University

Language Technologies Institute

Introduction

As an active NLP domain, emotion detection aims to determine the emotion(s) behind the plain text. GoEmotions dataset presents a multi-label classification challenge with more fine-grained emotion taxonomy than previous benchmarks. We propose a label-aware attention mechanism which significantly outperforms the baseline model.

Task

GoEmotions consists of 58K English Reddit comments, each labelled with one or more of the 28 emotions, including neutral. The baseline classifier proposed by the author is trained by finetuning a BERT model on this dataset, which yields the macro F1-score 0.46.

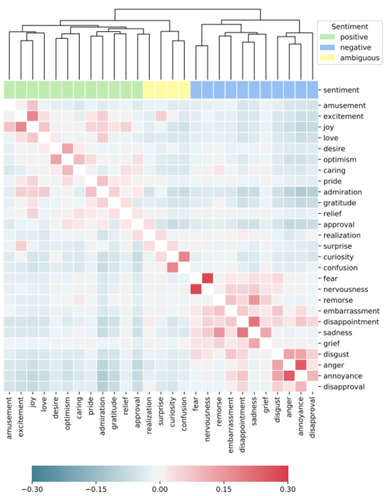

Figure 1:Label correlation

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 1:Statistics of the GoEmotions dataset. We see that 83% of the samples have single label, and class “neutral" makes up 26% of all samples.

The challenge of the GoEmotions dataset comes from its fine-grained taxonomy. As illustrated in Figure 1, many emotions are in fact semantically and statistically correlated. After performing error analysis on the baseline model, we found that “neutral" is often confused with other emotions, and low-resourced emotions are often misclassified as high-resourced ones with similar sentiment.

Method

To mitigate the naively assumed independence between emotions (labels), we introduce an label-aware attention mechanism, where each emotion is enabled to focus on different aspects of a sentence, depending on the semantic representation of the emotion label.

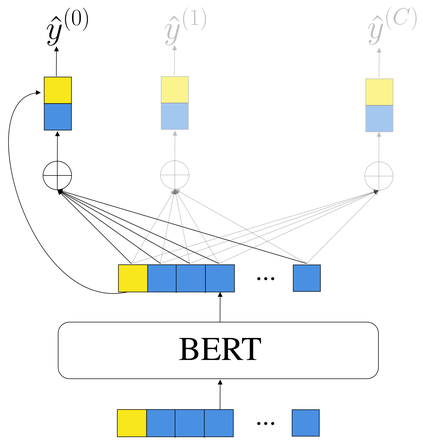

Figure 2:Visualization of the label-aware attention mechanism. The yellow block corresponds to the `[CLS]` token or its vector representation.

More formally, for an emotion i we create a trainable representation vector $\mathbf{e}^{(i)} \in \mathbb{R}^{d'}$ and use it to query the d-dimentional final-layer hidden states $\{\mathbf{h}_j\}_{j=1}^{L}$ from the pretrained language model, where L is the input sequence length. A `softmax` layer is applied to the bilinear product to compute the attention weights, which linearly combine contextualized embeddings:

$$\alpha_j^{(i)} = \frac{\exp(\mathbf{e}^{(i)T}\mathbf{W}_{\text{attn}}\mathbf{h}_j)}{\sum_{k=1}^{L}\exp(\mathbf{e}^{(i)T}\mathbf{W}_{\text{attn}}\mathbf{h}_k)} \qquad \tilde{\mathbf{h}}^{(i)} = \sum_{k=1}^{L}\alpha_k^{(i)}\mathbf{h}_k$$

where $\mathbf{W}_{\text{attn}} \in \mathbb{R}^{d' \times d}$ is a trainable transformation matrix and $\tilde{\mathbf{h}}_i$ is a label-awa re embedding for emotion i. To predict the likelihood of emotion i, we concatenate the sentence embedding $\mathbf{h}_0$ (which corresponds to the `[CLS]` token) and the label-aware embedding, and then pass the result to a dense layer with sigmoid activation:

$$\hat{y}^{(i)} = \sigma\left(\mathbf{w}_{\text{out}}^{T}[\mathbf{h}_0 \| \tilde{\mathbf{h}}^{(i)}] + b_{\text{out}}\right)$$

Results

Emotions	Baseline (F1)	Ours (F1)
Admiration	0.65	**0.68**
Amusement	**0.80**	**0.80**
Anger	0.47	**0.48**
Annoyance	0.34	**0.36**
Approval	0.36	**0.40**
Caring	0.39	**0.44**
Confusion	0.37	**0.42**
Curiosity	0.54	**0.55**
Desire	**0.49**	0.48
Disappointment	0.28	**0.33**
Disapproval	0.39	**0.41**
Disgust	0.45	**0.46**
Embarrassment	0.43	**0.45**
Excitement	0.34	**0.41**
Fear	0.60	**0.68**
Gratitude	0.86	**0.90**
Grief	0.00	**0.35**
Joy	0.51	**0.59**
Love	0.78	**0.80**
Nervousness	**0.35**	0.34
Neutral	**0.68**	0.66
Optimism	0.51	**0.55**
Pride	0.36	**0.50**
Realization	0.21	**0.25**
Relief	0.15	**0.35**
Remorse	0.66	**0.67**
Sadness	0.49	**0.53**
Surprise	0.50	**0.55**
Macro Avg.	0.46	**0.51**
Std	0.19	**0.01**

Table 2:F1 scores across all emotions in GoEmotions dataset, in comparison with baseline. Our model achieves significant improvement on Macro-average F1 and consistent improvements across all emotion groups. Moreover, the model shows much lower variance under different seeds, an indication of robust design.