11711: Reimplementing GoEmotions Baseline for Emotion Classification

Ethan Wu
yongyiw@andrew.cmu.edu

Yilin Wang
yilinwan@andrew.cmu.edu

Ziqi Liu
ziqil2@andrew.cmu.edu

1 Introduction

Emotion recognition is a task essential for many tasks, from dialog systems construction to spam detection. Emotion classification, as a subbranch of emotion detection, focuses on assigning labels that could be easily understood by humans to given input texts.

Our project was inspired by the *GoEmotions* datasets (Demszky, 2020)[1]. The main goal of this dataset is to produce a corpus using a more fine-grained emotion taxonomy than current Ekman's taxonomy. This paper also released a baseline emotion prediction model by fine-tuning BERT on their proposed dataset, achieving a macro-average F1-score of .46 across GoEmotion's taxonomy, leaving much room for improvement. As the dataset is relatively new, the baseline BERT model is also the state-of-the-art (SOTA) model. In this paper, BERT model finetuned on the proposed dataset shows transferability to other emotion predictions target datasets listed in *an analysis of annotated corpora for emotion classification in text.* (Bostan & Klinger, 2018) [12]

We have made code and data related to the experiment reproduction available at: https://github.com/yongyi-wu/nlp-group.

2 Literature Review

2.1 Fundamental knowledge of the Emotion Recognition in Conversation (ERC) task

Emotion taxonomy: There are two types of emotion categorization in the field: categorical and dimensional. On the categorical front, Ekman's model divides emotions into 6 fundamental emotions (fear, disgust, surprise, joy, anger, sadness), while Plutchik's emotion wheels divide emotions into 8 fundamental types, with each being able to subdivide into subcategories. For dimensional approaches, emotion can be expressed continuously in two dimensions: valence (degree of positivity) and arousal (degree of intensity).

GoEmotions dataset extends this traditional taxonomy to 27 subcategories (excluding neutral), allowing models to learn more fine-grained distinction between emotions in the corpora.

Existing Datasets for emotion classification in text: In *An Analysis of Annotated Corpora for Emotion Classification in Text* (2018), Bostan and Klinger aggregated existing emotion classification corpora into a one framework with unified annotation schema and file format (In GoEmotions' paper, the author uses 9 datasets mentioned in this paper as target datasets for their transferring learning experiments). The nine datasets are Daily Dialog [2], Emotion Stimulus [3], EmoInt [9], Electoral Tweets [5], Affective Text [4], ISEAR [6], CrowdFlower [7], TEC [8], SSEC [10]. All of the nine datasets have Ekman's annotation available excepts SSEC and Electoral Tweets, which uses Plutchik's annotations.

2.2 Emotion Recognition in Conversation (ERC)

Emotion classification is the main task that the $GoEmotions$ dataset addresses. Emotion classification is a subtask of Emotion Recognition in Conversation (ERC), which involves dialogue data

(multi-speaker, multiple utterances), where the model is given an entire conversation and is required to predict emotion labels for each utterance. Our main task, emotion classification, does not require an entire conversation as the input, but rather predicts emotion labels for a single Reddit comment.

However, it is noteworthy that emotion classification and ERC share huge common ground, and many techniques used in ERC tasks can be adapted to emotion classification to improve model performance. Moreover, although the processed go-emotion dataset only contains sentence-level Reddit comments, the raw data do include links to the parent text. Therefore, it is possible to use GoEmotions in ERC settings, which will allow our model to use contextual information. In the rest of this section, we will examine literature that is relevant to emotion classification as well as to ERC.

2.3 Challenges in the fields

In 2019, the paper *Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances*[13] (Poria et al., 2019) surveys the challenge of emotion classification tasks and drawbacks in recent models and approaches. Here we listed a few relevant points associated with emotion classification.

Sarcasm in dialog Detection of sarcasm is important in emotion recognition, as literal interpretations of emotions will almost definitely fail when conversations involve sarcasm.

Emotional reasoning It is possible for each part of the text to contain different emotions. The model must be able to recognize emotions in different parts of the sentence and reason through them logically for it to understand sentence-level semantics, which is crucial for classification.

2.4 Drawbacks of current approaches and Contribution of GoEmotions

Drawback of emotion model Ekman's categorization is too simple and unable to ground complex emotions, Plutchik's emotion wheel model is hard for annotators to distinguish emotions are different intensities (e.g. rage and anger). Recent datasets such as DailyDialog only deploy Ekman's emotion annotations.

Contribution of GoEmotions GoEmotions addresses the prior drawback by employing a 27-fold categorization of emotions. By including more complicated and subtle emotions like "realization" and "curiosity", we force the model to achieve a higher level of natural language understanding (NLU). If the model performs well in classifying complicated emotion, it must be able to perform better under simpler emotion taxonomies, like Ekman's model.

2.5 Related datasets and State-Of-The-Art (SOTA) models in Emotion Classification

2.5.1 Dataset: Affect in Tweets

The SemEval-2018 Task 1: *Affect in Tweets* is a dataset that contains annotated Tweets in English, Arabic, and Spanish. The dataset was created for the following 5 tasks 1) emotion intensity regression, 2) emotion intensity ordinal classification, 3) valence regression 4) valence ordinal classification and 5) emotion classification. For the emotion classification task, each tweet is annotated with one of the 12 emotions (including neutral) by multiple annotators, and the annotations were aggregated so that each tweet may receive multiple labels.

SpanEmo The SOTA model on *Affect in Tweets*, achieving a multi-label accuracy of 0.601 and macro F1-score of 0.578 on the English dataset. Consider over-lap of emotions within sentences. Cast multi-label emotion prediction into a span prediction problem. Aid ER (Emotion Recognition) model to associate labels with words in a sentence. (Potential direction for A4. From single-label emotion prediction to multi-label emotion prediction.)

BERT+DK In *Improving Multi-label Emotion Classification by Integrating both General and Domain-specific Knowledge* (Ying et al., 2019), the authors first finetuned BERT on twitter sentences to learn general linguistic knowledge, and they utilized a Twitter-specific text preprocessor (*ekphrasis*, Baziotis et al., 2017) to include domain-specific knowledge into the representation.

2.5.2 Dataset: ROCStories

ROCStories contains 50K 5-sentence stories that capture the causal and commonsense relations between daily events. Rashkin et al. annotated a subset of ROCStories with one or more of the eight emotion labels for each story.

Semi-supervision Gaonkar et al. found that adding the pretrained label semantics embeddings as input and incorporating the learned correlation between emotions in the loss function improves the performance over the baseline BERT model. Augmenting this method and applying semi-supervised techniques on the unlabeled portion of ROCStories gives the SOTA model with F1-score being 65.88, which is 4.92 point higher than the baseline BERT model.

2.5.3 Dataset: DailyDialog

DailyDialog collects human dyadic (two-party dialog) daily conversation as data, and deploys Ekman's emotion model for emotion annotation. This dataset does not annotate the speaker's name for each utterance.

CESTa In *Contextualized Emotion Recognition in Conversation as Sequence Tagging* (Wang et al., 2020) [17], Wang et al. were the first to cast ERC task to sequence tagging task. They used CRF to model the emotional consistency within context to find the best tag sequence for the entire conversation. They remain the SOTA result on the DailyDialog dataset.

2.5.4 Dataset: IEMOCAP

IEMOCAP is a multi-modal, dyadic (two-party) dialogue dataset, collected from 10 participant's conversations. IEMOCAP has about 5 times more utterances per dialogue than MELT. IEMOCAP doesn't assign speaker names

EmoBERTa In *EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa* (Kim & Vossen, 2021) [16] finetunes RoBERTa-large model on IEMOCAP and MELD datasets and achieves SOTA results respectively. Their creativity lies at the appending future utterances to the inputs into the RoBERTa model (three-segmented input, past utterances appending current utterances appending future utterances), and prepending speaker's name before each utterance, so that RoBERTa becomes speaker-aware, and hence able to model self and inter-personal emotional influences.

2.5.5 MELD

MELD is a multi-model, multi-party dataset derived from dialogues in TV series *Friends*. MELD's emotional class distribution is highly imbalanced, and hence a lot of model uses the weight-F1 score to measure model performance when trained on this dataset. MELD assigns speakers' names to each utterance.

TODKAT In *Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection* (Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, Yulan He), the author proposes the first topic-driven approach to the ERC task. They use an augmented LM for topic detection, and retrieve commonsense the knowledge from knowledge base based on conversational context. The fused representation and then fed into a Transformer encoder-decoder model to derive the emotion labels for every utterance.

3 Reimplementation Details

In our assignment 3 project, we rerun the fine-tune training of the paper's model (BERT base) on GoEmotions dataset, and produce similar scores. Based on Table 1, 2, and 3, our results (precision, recall, F1) are within the standard deviation of the reference results. In fact, our results differ from the reference by a very small amount, which is much less than the std.

We also rerun the transfer learning experiments in section 6 of the original paper and redraw the 9 plots of their results. See results in Figure 1.

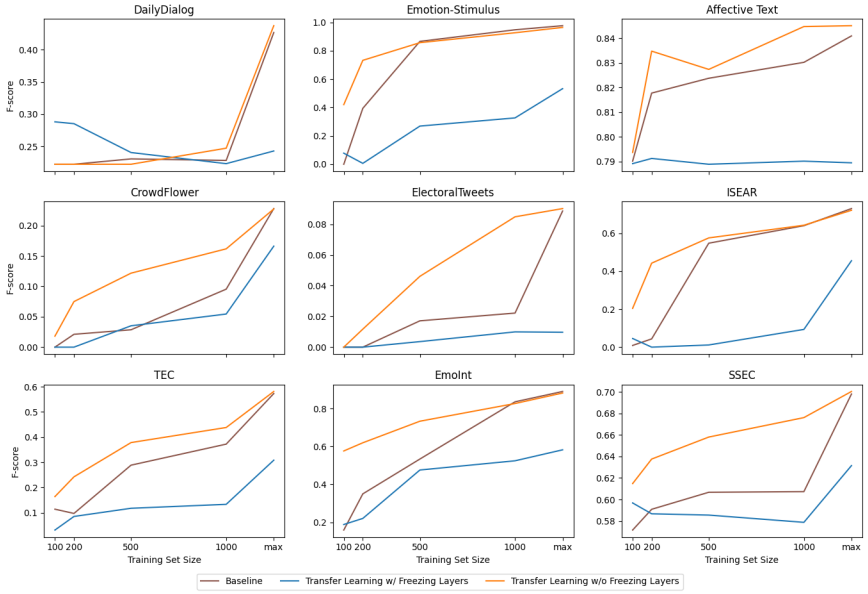

Figure 1: Transfer learning performance on 9 datasets

Table 1: Results based on GoEmotions taxonomy

Emotion	Precision	Recall	F1
admiration	0.633	0.738	0.681
amusement	0.750	0.909	0.822
anger	0.493	0.525	0.509
annoyance	0.342	0.431	0.381
approval	0.398	0.459	0.426
caring	0.440	0.459	0.449
confusion	0.362	0.516	0.426
curiosity	0.490	0.658	0.562
desire	0.583	0.506	0.542
disappointment	0.331	0.351	0.341
disapproval	0.366	0.431	0.396
disgust	0.483	0.463	0.473
embarrassment	0.600	0.405	0.484
excitement	0.429	0.466	0.447
fear	0.588	0.769	0.667
gratitude	0.920	0.912	0.916
grief	0.000	0.000	0.000
joy	0.539	0.646	0.588
love	0.745	0.870	0.802
nervousness	0.381	0.348	0.364
optimism	0.552	0.602	0.576
pride	0.667	0.125	0.211
realization	0.246	0.200	0.221
relief	0.000	0.000	0.000
remorse	0.584	0.804	0.677
sadness	0.549	0.571	0.560
surprise	0.545	0.553	0.549
neutral	0.630	0.672	0.651
macro-average	**0.487**	**0.514**	**0.490**
std (ref)	**0.18**	**0.24**	**0.19**

Table 2: Results based on sentiment-grouped data.

Emotion	Precision	Recall	F1
ambiguous	0.511	0.645	0.570
negative	0.639	0.727	0.680
positive	0.765	0.855	0.807
macro-average	**0.639**	**0.723**	**0.678**
std (ref)	**0.09**	**0.10**	**0.09**

Table 3: Results based on Ekman taxonomy.

Emotion	Precision	Recall	F1
anger	0.502	0.574	0.536
disgust	0.470	0.512	0.490
fear	0.612	0.724	0.664
joy	0.783	0.860	0.819
sadness	0.556	0.604	0.579
surprise	0.523	0.662	0.584
macro-average	**0.583**	**0.658**	**0.618**
std (ref)	**0.10**	**0.11**	**0.10**

4 Error analysis

In Table 2, we see that our model performs the best on positive emotions and worst on ambiguous emotions. More specifically, Table 1 shows that our model performs particularly badly in classifying emotions such as “realization", “pride", and “annoyance". A plausible explanation is that these emotions are more complicated and require more contextual information to properly classify.

As described above, GoEmotions characterizes a multi-label classification task. Our model performs this task by producing a probability measure for different labels for each sample. We use an aggregation procedure so that our prediction retains at most 3 labels for each sample.

For the sake of simplicity, we use the following procedures in later error analysis. If the sample contains only has 1 true label, then it is considered correctly classified if the top 1 predicted label is the same as the true label. If the sample contains multiple true labels, it is considered correct if all its label is contained in the top 3 predicted label.

Table 4 and 5 provide some most common mistakes made by our model using the GoEmotions classifications of emotion. We see in table 4 that the most common errors for single-label samples can be generalized as the misclassification between “neutral" and positive/negative emotion labels. We see in table 5 that for samples with multiple labels, our model often succeed in predicting one of the labels (which refers to simpler emotions such as “neutral" or “love"), but fail to recognize more complicated emotions (like “realization" and “annoyance").

Table 4: Top 7 common error for samples with only 1 true label

True Label	Predicted Label	Occurrence
neutral	approval	89
neutral	disapproval	72
neutral	curiosity	68
approval	neutral	67
neutral	annoyance	61
disapproval	neutral	56
annoyance	neutral	55

In table 6 we present 2 typical misclassified examples that are perhaps indicative of how the error is made. It is plausible to argue that the classifier did not achieve a high understanding of the semantics of the sentence, but it utilizes only certain characteristics of the sentence to make the classification. For example, the reason that sentence (A) is classified as “annoyance" is probably because it contains the word “moron", which is usually indicative of annoyance or anger. Similarly, the reason that sentence (B) is classified as “curious" is probably that contains“?", which is usually indicative of curiosity. However, in (A) and (B), the presence of these two words does not mean that the sentence exhibit other emotion other than “neutral".

Therefore, it is plausible to say that the BERT model did not achieve a high level of natural language understanding (NLU). The use of a more advanced NLU model that incorporates semantics is necessary to improve the model performance.

Table 5: Top 5 common error for samples with multiple true label

True Label 1	True Label 2	Predicted Label	Occurrence
annoyance	neutral	neutral	10
approval	neutral	neutral	8
admiration	love	love	7
approval	admiration	admiration	7
realization	neutral	neutral	6
annoyance	disapproval	neutral	5

Table 6: Two representative mistakes

ID	Sentence	True Label	Predicted Label
(A)	That's a very long way of saying "I'm a moron"	anger	annoyance
(B)	You know that mista mista lady? I think I just killed her. - [NAME]	neutral	curiosity

5 Further Improvements

In section 4, we argued that our model did not achieve a high level of natural language understanding and we need to use more advanced NLU models. To address this issue, we could use some SOTA NLU models such as HNN (He et al.) [19] in future work.

In Section 4, we also see that our model often fails to distinguish "neutral" from other emotions. To address this issue, it may be helpful to use contrastive learning in our model to force the model to separate (embedding of words) "neutral" from other emotions. For example, in finetuning, we could include contrastive learning objectives in the loss function, as suggested by Gunel et al. (2021) [21]. It is also possible to adapt existing pretrained language models that were trained on contrastive learning objectives, such as CERT (Fang et al., 2020) [20].

The work by Gaonkar et al., which highlights the importance of the label semantics, also appears to be a promising direction in the emotion classification task. This is because incorporating label semantics enables us to avoid treating each category independently but leverage inherent distance within and across positive, neutral ,and negative emotions. Therefore, we can consider replacing the label token with label embedding, including the label embedding as inputs, and taking label correlation into account in the loss function. These modifications can potentially address the aforementioned vulnerabilities of the baseline model.

References

[1] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi (2020). *GoEmotions: A Dataset of Fine-Grained Emotions*, arXiv:2005.00547 [cs.CL].

[2] Li Yanran, Su Hui, Shen Xiaoyu, Li Wenjie, Cao Ziqiang, and Niu Shuzi (2017). *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*, Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing.

[3] 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.

[4] Carlo Strapparava and Rada Mihalcea (2007). *SemEval2007 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.

[5] 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.

[6] 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.

[7] CrowdFlower (2016). https://www.figureeight.com/data/sentiment-analysis-emotion-text/.

[8] 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.

[9] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko (2018). *SemEval2018 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.

[10] Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Pado, 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.

[11] 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).

[12] 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.

[13] Soujanya Poria, Navonil Majumder, Rada Mihalcea, Eduard Hovy (2019). *Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances*. arXiv:1905.02947.

[14] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann (2018). *Conversational memory network for emotion recognition in dyadic dialogue videos*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), vol. 1, 2018, pp. 2122–2132.

[15] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann (2018). *Icon: Interactive conversational memory network for multimodal emotion detection*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.

[16] Taewoon Kim, Piek Vossen (2021). *EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa*. arXiv:2108.12009.

[17] Wang Yan, Zhang Jiayu, Ma Jun, Wang Shaojun, Xiao Jing (2020). *Contextualized Emotion Recognition in Conversation as Sequence Tagging*. Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics.

[18] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, Yulan He (2021). *Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection*. arXiv:2106.01071.

[19] Pengcheng He, Xiaodong Liu, Weizhu Chen, Jianfeng Gao (2019). *A Hybrid Neural Network Model for Commonsense Reasoning*. arXiv:1907.11983.

[20] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, Pengtao Xie (2020). *CERT: Contrastive Self-supervised Learning for Language Understanding*. arXiv:2005.12766.

[21] Beliz Gunel, Jingfei Du, Alexis Conneau, Ves Stoyanov (2021). *Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning.* arXiv:2011.01403.