An Ensemble Approach to Detect Emotions at an Essay Level

Himanshu Maheshwari[1] and **Vasudeva Varma** [2]
IIIT Hyderabad
[1] himanshu.maheshwari@research.iiit.ac.in, [2] vv@iiit.ac.in

Abstract

This paper describes our system (IREL, reffered as himanshu.1007 on Codalab) for Shared Task on Empathy Detection, Emotion Classification, and Personality Detection at 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis at ACL 2022. We participated in track 2 for predicting emotion at the essay level. We propose an ensemble approach that leverages the linguistic knowledge of the RoBERTa, BART-large, and RoBERTa model finetuned on the GoEmotions dataset. Each brings in its unique advantage, as we discuss in the paper. Our proposed system achieved a Macro F1 score of 0.585 and ranked one out of thirteen teams (the current top team on leaderboard submitted after the deadline). The code can be found here

1 Introduction

Emotion is a concept that is challenging to describe. Nevertheless, as human beings, we understand the emotional effect situations have or could have on other people and us. In this work, we aim to transfer this knowledge of emotion detection to machines. This work aims to develop a robust system that could detect emotions at an essay level. These essays are reactions to news stories and are between 300 and 800 characters in length.

Existing literature on emotion detection mainly focuses on emotion detection at the sentence level. Different datasets consisting of sentences from social media (Mohammad (2012), Mohammad et al. (2014), Liu et al. (2017), Demszky et al. (2020)), fairytales (Alm and Sproat, 2005), dialogues (Li et al., 2017), etc. have been made available. However, the task of emotion detection at an essay level is underexplored. In essay-level emotion detection, the emotions are typically expressed by the entire narrative and not just a few words or phrases. The system must refer to the entire essay to get a more holistic view of the expressed emotion. We empirically show that systems trained on just sentence level emotion detection will not work essay level as they do not have the entire context.

We propose an ensemble approach consisting of a finetuned RoBERTa (Liu et al., 2019), finetuned BART-large (Lewis et al., 2020), and RoBERTa model first finetuned on the GoEmotions (Demszky et al., 2020) dataset and then finetuned on our dataset. RoBERTa model has shown amazing performance for various NLP tasks and thus was the default choice for the task. BART-large has shown amazing performance for summarization tasks. This suggests it is suitable for a task involving multiple sentences. The last model is a RoBERTa model that was first finetuned on the GoEmotions dataset and then finetuned on our dataset. The intuition is that since it has a good understanding of sentence-level emotions (from GoEmotions), it will combine the sentence-level knowledge into essay-level knowledge. This is especially important for cases with very strong expression of emotions in a sentence. Ablation studies show that the model performs worse in the absence of either of the three models. Another ablation study is conducted to reinforce our claim that the task can't be solved by looking at sentence level.

2 Dataset

The training dataset is a small supervised dataset consisting of various fields. However, only two fields are helpful for emotion prediction: essay and emotion; thus, we use only these fields. The dataset statistics are shown in table 1 and table 2.

The dataset is very small and heavily skewed, with anger and sadness making up ~54% of the entire dataset. This skewed dataset affects the model's performance, and it needs to be dealt with.

Usually, NLP systems deal with skewed datasets using oversampling, undersampling, augmentation, or weighted loss function. With such few data points, oversampling and undersampling are not

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 276 - 279

Split	Number of samples
Train	1860
Dev	270
Test	525

Table 1: Dataset statistics for different splits.

Emotion	Number of samples
Anger	349
Disgust	149
Fear	194
Joy	82
Neutral	275
Sadness	647
Surprise	164

Table 2: Emotion distribution for train set.

viable. Our initial exploration with data augmentation did not help; thus, we used a weighted cross-entropy loss function to deal with data imbalance. The weights of each class were determined using the sklearn library. [1]

3 Baselines

The following section describes different approaches we tried before shifting to our proposed methodology. For each approach, grid search was used to find appropriate hyperparameters. Please note we compare different models using Macro F1 score which is the official evaluation metric.

3.1 Language Model Finetuning

The current de facto in NLP is to finetune a language model for any classification task. Our first approach was to finetune a language model and observe the results. This will serve as a baseline for other approaches. This exercise also helps us select the appropriate language model for other approaches. We experimented with the following language models:

1. Roberta Base
2. Bert-base-uncased
3. Roberta-large
4. Bart-large
5. Longformer-base-4096

Table 4 shows the results of different language models. Roberta-base is performing the best; thus, it is the suitable language model for other approaches. Roberta-large overfits and was producing the same results after each epoch. Longformer, though suitable for long sequences, did not perform well.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

Emotion	Number of samples
Anger	2000
Disgust	2000
Fear	1230
Joy	2000
Neutral	2000
Sadness	2000
Surprise	2000

Table 3: Emotion distribution for train set of GoEmotions Dataset

3.2 Binary Classifiers

Having a classifier doing multiclass classification is challenging. In this approach, we use a binary classifier for each emotion and take the emotion with the highest softmax classification probability. Specifically, we finetune a Roberta-base binary classifier for each emotion. The classifier aims to identify target emotion from other emotions. During inference, we take the classification probability from each classifier. The emotion with the highest classification probability from its classifier is the predicted emotion. Table 4 shows the result of this approach. The results are poor compared to finetuning a classifier; thus, a binary view of emotion is unsuitable for our use case.

3.3 Finetuning a classifier trained on GoEmotions dataset

This approach introduces an additional layer of transfer learning. We first finetune a Roberta-base model on a subset of the GoEmotions dataset. GoEmotions is a sentence-level fine-grained emotion classification dataset. We take sentences that have only one of the seven emotions of our task. This GoEmotions finetuned classifier is then further finetuned on our dataset. The idea is to finetune a classifier that has some understanding of emotions. Table 3 shows statistics of the GoEmotions dataset. Table 4 shows results for the same. The results are poor, suggesting that strong sentence-level understanding does not scale to essay-level understanding.

Model	Macro F1	Accuracy in %
Finetuning Roberta Base	0.6090	**70.000**
Finetuning Bert base uncased	0.5502	62.593
Finetuning Roberta Large	0.0760	36.296
Finetuning BART Large	0.5983	66.667
Finetuning longformer-base-4096	0.5635	66.667
Combining Binary Classifiers	0.4689	63.333
Finetuning model trained on GoEmotion Dataset	0.5568	63.333
Proposed Solution	**0.6360**	68.519
Proposed Solution	**0.6360**	68.519
Proposed Solution w/o Roberta	0.6021	67.037
Proposed Solution w/o BART	0.6067	**69.259**
Proposed Solution w/o GoEmotions Roberta	0.6248	67.778
Roberta-base with entire sequence	**0.6090**	**70.000**
Roberta-base with sentence seperated sequence	0.5812	65.185

Table 4: Results of different models on dev set.

4 Proposed Approach

We make the following observations from the baseline models:

a. Roberta-base and Bart-large perform better than the rest of the language model. Both models bring their advantage, Roberta-base is a powerful language model for NLU tasks, and Bart-large is suitable for tasks involving multiple sentences.

b. Roberta-base model that is first finetuned on GoEmotions dataset followed by finetuning on our dataset performs poorly compared to other baselines. However, it has a firm sentence-level understanding. Thus, this model is suitable for samples with very strong emotional sentences.

Based on these observations, we combine the strength of the Roberta-base, Bart-Large, and Roberta-base model that is first finetuned on the GoEmotions dataset in an ensemble fashion. More specifically, we take the linear combination of classification probability by each model and predict the emotion with the highest classification probability (or score). Thus the classification probability (or score) is given by:

$$s_{emo} = \lambda_1 P_{RB} + \lambda_2 P_{BL} + \lambda_3 P_{RBG}$$

Where s_{emo} is the classification score for a particular emotion and λ_1, λ_2, λ_3 are the weights of each model. P_{RB} is the classification probability of Roberta-base, P_{BL} is the classification probability of BART large and P_{RBG} is the classification probability of Roberta-base finetuned on GoEmotions. The emotion with the highest score is predicted. We found λ_1, λ_2, and λ_3 using grid search on the dev set. The value that gave the best result is λ_1: 0.26, λ_2: 0.26, and λ_3: 0.07. Table 4 shows the results of this approach. This approach outperforms all the baselines on the dev set, suggesting strength in using multiple language models.

5 Training

As discussed, we use grid search to find the appropriate hyperparameters. We use a batch size four and a dropout of 0.3 for Roberta-base. For Bart-Large, we use a batch size of three and a dropout of 0.4. For Roberta-base trained on the GoEmotion dataset, we use batch size eight and dropout of 0.2 for the first layer of finetuning. For the second layer of finetuning, we use a batch size four and a dropout of 0.3. The learning rate and seed were fixed to 10^{-5} and 42, respectively. The training was done on Nvidia RTX 2080 TI (11 GB) and took about one hour for each model finetuning.

6 Results

Table 4 shows the results of our dev set. We submitted the ensemble solution discussed above based on hyperparameters and results on the dev set. Table 5 shows the test set results as reported on the Codalab platform. The proposed system achieved rank two.

7 Ablation Studies

We conducted two ablation studies to better understand our proposed approach and the problem setting.

Metric	Result
Macro F1-Score	0.585
Micro F1-Score	0.661
Accuracy	0.661
Macro Precision	0.594
Macro Recall	0.584
Micro Precision	0.661
Micro Recall	0.661

Table 5: Results on test set as reported on Codalab

7.1 Role of Each Language Model

In the first ablation study, we inspect the role of each language model described in the ensemble solution. We observe the performance by removing one model at a time. Table 4 shows the results for the same. We see that removing even one language model degrades the overall performance. This builds confidence in our choice and intuition behind each language model for the ensemble solution, and each of the three language models is essential for our task.

7.2 Sentence Level Treatment of the Task

This ablation study inspects the model's performance if we treat the input at a sentence level. Specifically, instead of inputting the entire essay to the Roberta-base, we input the essay separated into individual sentences. We break the essay into sentences and separate them using a special token used in Roberta-base to separate sequences. Table 4 shows the result of this ablation study. For a fair comparison, we compare results between a Roberta-base model fed the entire sequence, and a Roberta-base model fed the sentence separated sequence. We see that a Roberta-base model that is fed the entire sequence performs better than a Roberta-base model that is fed a sentence-separated sequence. This suggests that we need to look at the entire sequence for a holistic understanding of the emotion, and we cannot just rely on sentence-level information.

8 Conclusion

In this work, we explore the task of emotion prediction at an essay level. We first explore different language models and identify Roberta-base and Bart-large suitable for the task. Next, we observe that adding an additional layer of transfer learning by finetuning on a sentence-level dataset helps identify essays with very strong emotional sentences. Building on these two hypotheses, we propose an ensemble solution that combines the linguistic knowledge of Roberta-base, Bart-large and Roberta-base finetuned on the GoEmotions dataset. Our proposed solution achieved a macro F1 score of 0.585 and was ranked one globally (the current top team on leaderboard submitted after the deadline).

References

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In *Affective Computing and Intelligent Interaction*, pages 668–674, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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. *CoRR*, abs/1710.03957.

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. pages 477–483.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: 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 (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2014. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing Management*, 51.