WASSA@IITK at WASSA 2021: Multi-task Learning and Transformer Finetuning for Emotion Classification and Empathy Prediction

Jay Mundra[*] **Rohan Gupta**[*] **Sagnik Mukherjee**[*]
Indian Institute of Technology Kanpur (IIT Kanpur)
{jaym, rohangpt, sagnikm}@iitk.ac.in

Abstract

This paper describes our contribution to the WASSA 2021 shared task on Empathy Prediction and Emotion Classification. The broad goal of this task was to model an empathy score, a distress score and the overall level of emotion of an essay written in response to a newspaper article associated with harm to someone. We have used the ELECTRA model abundantly and also advanced deep learning approaches like multi-task learning. Additionally, we also leveraged standard machine learning techniques like ensembling. Our system achieves a Pearson Correlation Coefficient of **0.533** on sub-task I and a macro F1 score of **0.5528** on sub-task II. We ranked **1**st in Emotion Classification sub-task and **3**rd in Empathy Prediction sub-task.

1 Introduction

With the growing interest in human-computer interface, machines still lag in having and understanding emotions. Emotions are considered a trait of a living being and are often used to list differences between machines and living beings. Human emotions such as empathy are hard to describe even for humans, and a consensus is hard to be reached. The inherent knowledge humans have for these emotions is hard to pass on to machines, and hence this task is challenging. That is also probably the reason why the prior work in this area is really limited. Although there has been some research done by Xiao et al. (2012), Gibson et al. (2015) and Khanpour et al. (2017), they have some significant limitations, such as the oversimplified notion of empathy and unavailability of these corpora in the public domain. Sedoc et al. (2020) investigated the utility of Mixed-Level Feed Forward Network and also created an empathy lexicon.

Another problem in the affect domain is that of emotion classification of textual data. Klinger et al. (2018) presented an analysis of many datasets for this task. These datasets vary in size, origin and taxonomy. The most frequent emotion taxonomy used is that proposed by Ekman (1992) - identifying anger, disgust, fear, joy, sadness and surprise as the 6 emotion categories. Demszky et al. (2020) presented a dataset obtained from Reddit comments and tagged according to 27 fine-grained emotion categories identified by them. They also present a baseline obtained by fine-tuning a BERT-base (Devlin et al., 2019) model on their dataset.

In the WASSA 2021 Shared Task (Tafreshi et al., 2021) there are two major sub-tasks addressing the issues of empathy and distress prediction and emotion classification.

Track I: Empathy Prediction - the objective is to model the empathy concern as well as the personal distress at the essay level. This is a regression task.

Track II: Emotion Prediction - the objective is to predict the overall emotion of the essay. The labels are categorical, and it is a classification task.

In our approach, we have used transformer-based language models, primarily ELECTRA (Clark et al., 2020). We also experimented across different forms of multi-task learning, keeping ELECTRA at the base and having several feedforward layers on top of it, responsible for different tasks. It was also observed that various ensembling techniques worked pretty well for the tasks and improved the scores due to a better bias-variance trade-off.

2 Dataset

For this task, an extended version of the dataset by Buechel et al. (2018) was used, as per the official release from the task organizers.

The dataset contained 1860 data points for training where each data point was a tuple of the following - the essay, an empathy score, a distress score, age

[*] Authors equally contributed to this work.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 112–116
April 19, 2021. ©2021 Association for Computational Linguistics

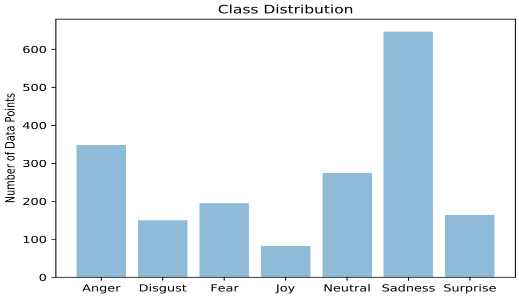


Figure 1: Frequency of classes in the dataset

and income of the annotator (demographic factors), overall emotion of the essay and various metrics denoting the personality of the respondent. The empathy and distress scores were in the range $(1, 7)$, and they were annotated with a 7-point scale. The validation and test sets had 270 and 525 data points, respectively. The emotion classes used in the task are closely related to the emotion classes discussed in Ekman (1992).

For this task, the volume of data was insufficient for finetuning large language models like ELECTRA. To overcome this, we experimented with data augmentation; in particular, we used the Ekman grouped data for emotion classification provided in the GoEmotions dataset (Demszky et al., 2020). The pros and cons of this methodology will be discussed in Section 3.2.2

3 System Description

3.1 Empathy Prediction

We propose two separate approaches for this task.

3.1.1 Transformer finetuning

The first approach is based on finetuning an ELECTRA large model separately for empathy and distress with the Mean Squared Error (MSE) loss function. The ELECTRA parameters were kept trainable, and it was finetuned with a single feed-forward layer on top of it. We experimented with the number of linear layers and the number of hidden neurons and chose a single layer since it yielded the best validation score. We shall refer to this approach as the Vanilla ELECTRA method.

3.1.2 Finetuning with Multi-task learning

Multi-task learning has recently caught a lot of interest in the NLP community (Liu et al. (2019a), Worsham and Kalita (2020)). The objectives, namely empathy and personal distress, are seemingly closely related, and hence a multi-task learning setup was used, jointly modelling them. We used an ELECTRA-large with two dense layers on top of it, one responsible for Empathy and another for Distress. Like the previous approach, MSE loss was used, adding the loss for Empathy and Distress and jointly training the architecture end to end on that total loss.

The same approach was tried out with the RoBERTa (Liu et al., 2019b) model.

3.1.3 Final Ensemble

For the empathy prediction task, the final system was an ensemble of two models - Roberta multi-task model and Vanilla ELECTRA model.

For the distress prediction task, the final system was an ensemble of two models, both being ELECTRA models trained with multi-task loss, with different performances on the development set. We finally submitted an ensemble of two models for each task - Empathy prediction, and Distress prediction.

For all these ensembling, a simple average of the output across the trained models was taken.

3.2 Emotion Classification

3.2.1 Transformer finetuning

In this approach the ELECTRA model was finetuned. The [CLS] token was passed through a single linear layer to produce a vector of size 7, representing class probabilities (or scores). We trained using the cross entropy loss function.

As per our observations these models were sensitive to initialisation. The validation accuracy scores varied significantly across different seed values. Hence, the models were trained several times, and the snapshots with best validation scores were saved.

3.2.2 Data Augmentation

The Figure 1 clearly shows that there was a heavy class imbalance in the dataset provided by the organizers, ‘joy’ being the least represented class (since the data collected is related to harm done to someone). Class imbalance is a standard problem faced by the machine learning community in classification problems (Longadge and Dongre (2013), Haixiang et al. (2017)). To address this issue, the dataset was augmented with the GoEmotions dataset (Demszky et al., 2020), since the class labels of the Ekman variant of it was exactly the same. However, this dataset was different from ours because the

Method	Val Macro F1
+ aug	0.6042
- aug	0.561

Table 1: Variation of validation macro F1 scores with and without the data augmentation technique for classification task for the ELECTRA base model ('+ aug' means data augmentation has been used and '- aug' means otherwise.)

essays' length was significantly lesser for GoEmotions dataset than our task dataset. Hence the appropriate number of data points to be sampled was an important hyperparameter. Too much sampling would make the distribution of the train data very different from the distribution of validation and test data, and would cause the model to eventually fail to generalise.

The augmentation scheme was different across different components of the final ensemble. While an ELECTRA base was finetuned on a class balanced dataset of total 2800 samples represented by BA (Balanced Augmentation) in Table 2, another ELECTRA base was finetuned with randomly chosen 1000 samples represented by RA (Random Augmentation) in Table 2. Same augmentation scheme was followed on ELECTRA Large and RoBERTa. We also trained an ALBERT (Lan et al., 2019) model on the BA augmented data. The intuition behind using multiple transformer models was to have different "strength" of different models.

3.2.3 Final Ensemble

We created two ensemble models for this task by summing the probability scores of the models involved in ensemble. The final systems submitted for this task were two ensembles. The ensembling technique was to take sum or average across class scores and then selecting the class with the highest score. One of them (Ensemble 1 in table 2) was an ensemble of two ELECTRA base models and one ELECTRA large model (model 1, 2, 3 in the table 2) that were trained on different augmentation schemes.

The second ensemble (Ensemble 2 in table 2) is the ensemble of the first 7 models shown in Table 2 comprising of 2 ELECTRA base, 2 ELECTRA large, 2 RoBERTa, and an ALBERT trained using the methods outlined in the previous subsection.

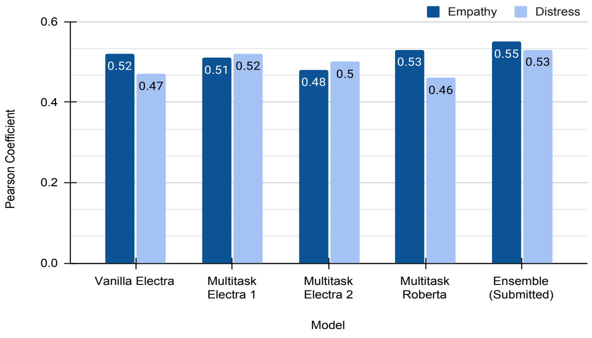


Figure 2: Pearson's Correlation Coefficient of Sub-task I on dev data

Model	Macro F1	Accuracy
ELECTRA base (RA)	0.604	**69.25**
ELECTRA base (BA)	**0.608**	67.77
ELECTRA Large (RA)	0.582	68.51
ELECTRA Large (BA)	0.585	66.29
RoBERTa Large (RA)	0.588	67.03
RoBERTa Large (BA)	0.583	66.29
ALBERT Large (BA)	0.595	68.51
Ensemble 1*	0.64	71.11
Ensemble 2	**0.65**	**72.59**

Table 2: Accuracy and macro-F1 scores on emotion classification on the dev data. * - Submitted model

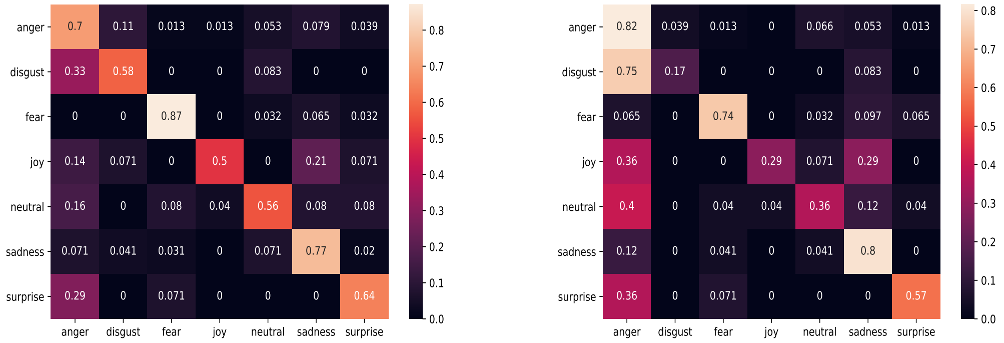
4 Experimental Setup

For all tasks the learning rate was set to 10^{-5}, and the models were trained using AdamW (Loshchilov and Hutter, 2017) as optimizer. The parameters for AdamW were $\beta = (0.9, 0.99)$, $\epsilon = 10^{-6}$, weight_decay = 0. Batch size was set to 16 for the Section 3.1.1 approach, and set to 8 for Section 3.1.2 and Section 3.2.1 with the shuffle parameter set to True on pytorch dataloader. A single Tesla V100-SXM2-16GB GPU was used for the finetuning experiments. The GPUs were available on the Google Colab platform.

The ELECTRA and RoBERTa were used off the shelf from the HuggingFace library (Wolf et al., 2020).

5 Results

The results on the empathy prediction sub-task and the distress prediction sub-task on the development set can be found in Figure 2. Also presented in the figure is the performance of the ensemble models that we submitted for evaluation. On the test data, the ensembles achieved a pearson correlation



(a) Final Submission

(b) Electra base without augmentation

Figure 3: Confusion matrix (normalized) for the emotion classification task on the dev data

coefficient of **0.558** and **0.507** on the Empathy prediction, and Distress prediction respectively. This amounts to an average score of **0.533** which ranks us 3^{rd} in this sub-task.

In Table 1, we present a result on the development set with and without using balanced data augmentation (BA) using GoEmotions for the ELECTRA base model. As we expected, the data augmentation helped improve performance of the model. The results for the various EMO models on the development set are available in Table 2. We also list the performance of two final ensembles in the table. On the test data, the Ensemble 1 and Ensemble 2 achieved a macro F1 score of **0.5528** and **0.588** respectively. We submitted Ensemble 1 as our final submission in the evaluation phase, as we were allowed only one submission. We ranked 1^{st} in this sub-task.

6 Error Analysis

It was observed that the training was highly sensitive to the initialization of the models. These include the initializations of the weight vectors of the feed-forward layers, and the ordering and organization of the batches fed to the model. Across different seeds the models' scores varied significantly. This is in line with the analysis done by Dodge et al. (2020) for transformer based models on the GLUE Benchmark (Wang et al., 2018). The high variability in performance can be observed Fig 2 . The Multi-task ELECTRA model performed differently on different runs, and we list two such runs for each Empathy Prediction and Distress Prediction.

The confusion matrix in Figure 3 shows that while the submitted system performed extremely well on the fear class, it underperformed a bit on joy and neutral, the performance on joy being very low. A point worth noting might be that the only 'positive' human emotion here is joy, and all the others are negative emotions and are often hard to distinguish by humans. A fare share of data points have been labeled as sadness and anger while they actually belong to the class joy. We also present, for comparison, the confusion matrix of the ELECTRA base model trained on non-augmented data. This model performed much worse on all emotions except sadness and anger; the model being very prone to predict the emotion as anger. The good performance of this model on sadness could be because of high number of samples from the that class in training data. The tendency to predict anger correlates with the fact that anger is the 2nd most frequent label in the training data. From this comparison, we can say data augmentation helped us achieve a more balanced performance with respect to all emotions, in particular bringing down the tendency to predict anger as the emotion, and improving performance for all other emotion classes which had less training data.

7 Conclusion

This paper describes our submission to the WASSA 2021 shared task, where we have leveraged off the shelf transformer models pre-trained on huge corpora in the English language. The intuition for keeping these models was to exploit the huge information they already possess. In the evaluation phase our systems ranked 1st in track II and 3rd in track I.

References

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

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

James Gibson, Nikos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S. Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *INTERSPEECH*.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data. *Expert Syst. Appl.*, 73(C):220–239.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Roman Klinger et al. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

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

Rushi Longadge and Snehalata Dongre. 2013. Class imbalance problem in data mining review.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Joseph Worsham and Jugal Kalita. 2020. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters*, 136:120–126.

B. Xiao, D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4.