Transformer-based Architecture for Empathy Prediction and Emotion Classification

Himil Vasava **Pramegh Uikey** **Gaurav Wasnik** **Raksha Sharma**
Indian Institute of Technology Roorkee (IIT Roorkee)
{vasava_h, uikey_p, wasnik_g, raksha.sharma}@cs.iitr.ac.in

Abstract

This paper describes the contribution of team PHG to the WASSA 2022 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 type of emotion associated with the person who had reacted to the essay written in response to a newspaper article. We have used the RoBERTa model for training, and on top of it, five layers are added to finetune the transformer. We also use a few machine learning techniques to augment and upsample the data. Our system achieves a Pearson Correlation Coefficient of 0.488 on Task 1 (Average of Empathy - 0.470 and Distress - 0.506) and Macro F1-score of 0.531 on Task 2.

1 Introduction

Empathy and Distress are quite important regarding human health. Emotion classification in natural languages has been studied for over two decades, and many applications successfully used emotion as their principal component. Empathy utterances can be emotional. Therefore, examining emotion in text-based empathy has a significant impact on predicting empathy. Empathic concern and personal distress are empathic responses that may result when observing someone in discomfort (Fabi et al., 2019). Some news stories are also displayed in this task, and people have reacted to them. The news is disturbing or discomforting to some people. And hence, regarding that, their empathy and distress are noted. This paper presents the WASSA 2022 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. This shared task included four individual tasks where teams developed models to predict Emotions, empathy, and personality in essays in which people expressed their empathy and distress in reaction to news articles in which an individual or a group of people were harmed. Additionally, the dataset also included the demographic information of the authors of the essays, such as age, gender, ethnicity, income, education level, and personality information. The shared task consisted of four tracks (optional):
Track 1: Empathy Prediction (EMP) task consists of predicting both the empathy concern and the personal distress. (Evaluation based on an average of Pearson correlation (Benesty et al., 2009) of empathy and distress).
Track 2: Emotion Classification (EMO) consists of predicting the emotion (sadness, joy, disgust, surprise, anger, or fear, taken from the six basic emotions (Ekman and Friesen, 1971) also including neutral) at the essay-level (Evaluation based on the macro F1-score).
Track 3: Personality Prediction (PER), which consists in predicting the personality of the essay writer, knowing all their essays and the news article from which they reacted (Evaluation based on the average of Pearson correlation over Personality values (Komarraju et al., 2011) - conscientiousness, Openness, Extraversion, Agreeableness, and Stability).
Track 4: Interpersonal Reactivity Index Prediction (IRI) consists of predicting the personality of the essay writer. (Evaluation based on an average of Pearson correlation over IRI values - fantasy, perspective taking, empathetic concern, personal distress).
We participated in only the first two tasks.

2 Related Work

Over the last few years, earnest endeavors have been made in the NLP community to analyze empathy and distress. For text-based empathy prediction, (Buechel et al., 2018) laid a firm foundation for predicting Batson's (Batson et al., 1987) empathic concern and personal distress scores in reaction to news articles. They present the first publicly available gold-standard dataset for text-based empathy and distress prediction. To annotate emotions in text, classical studies in NLP suggest categorical

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 261 - 264

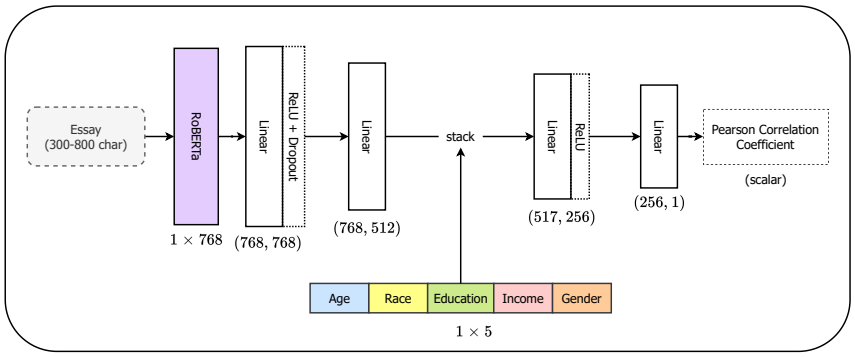

Figure 1: System Architecture

Set	Examples
Train	1860
Dev	270
Test	525

Table 1: Train-dev-test split

tagsets, and most studies are focussed on basic emotion models that psychological emotion models offer. The most popular one is the Ekman 6 basic emotions (Ekman and Friesen, 1971). The emotions presented in this dataset are the same six emotions by Ekman plus one extra emotion (neutral).

3 Dataset

The dataset is an extension to the one provided by (Buechel et al., 2018). For all the tasks, a train-dev-test split was provided. The dataset consists of essays collected from participants who had read news articles about a person, a group of people, or disturbing situations. The dataset had an essay (300-800 characters), empathy score, a distress score, emotion label, and other demographic information (age, gender, race, education, income) as well as personality information (conscientiousness, openness, extraversion, agreeableness, stability) and interpersonal reactivity index (IRI) scores (fantasy, perspective taking, empathetic concern, personal distress).

3.1 Data Augmentation

A single sentence does not always convey the information required to translate it into other languages; we sometimes need to specialize words that are ambiguous in the source languages (Sugiyama and Yoshinaga, 2019). So, we used back translation (Edunov et al., 2018) for text augmentation. The idea here was to have different sentences having the same meaning for training. Step 1: Select the essay (English).
Step 2: Select a random language and convert the essay to that language.
Step 3: Now translate that converted essay back to English.

We used Google translate API for translating essays back and forth. Every example was translated to one other language, and hence after back translation, the total number of samples was doubled (3720). Data augmentation improved the performance, as shown in the Table 2.

4 System Description

4.1 Empathy Prediction

Transformers (Vaswani et al., 2017) have outperformed recurrent neural networks (RNNs) in natural language generation (Kasai et al., 2021). For this task, we had to predict empathy and distress scores which had been done by training the same model by keeping the targets different (empathy for model 1 and distress for model 2). The approach used is based on fine-tuning RoBERTa model (Liu et al., 2019) separately for empathy and distress. To take the essay as input to the RoBERTa model, initially, tokenization (Webster and Kit, 1992) was required. The input tokens were made using the Roberta Tokenizer imported from the transformer library. The loss function used was Mean Squared Error (MSE). No parameters were frozen (all of them were trainable), and on top of it, five layers were trained (to make the network deeper). Four layers were linear, while one was a dropout layer (to prevent overfitting). In the pre-final layer, five additional demographic features were taken as input.

The model was trained on both the augmented

Metric	Original	Augmented
Macro F1-Score	0.5174	0.5311
Micro Recall	0.6152	0.6114
Micro Precision	0.6152	0.6114
Micro F1-Score	0.6152	0.6114
Macro Recall	0.5054	0.5288
Macro Precision	0.5461	0.557
Accuracy	0.6152	0.6114

Table 2: Original vs Augmented on Test set

data and original data. Still, the final submission was made using the model trained on the augmented data as it resulted in a higher Pearson Correlation Coefficient.

4.2 Emotion Classification

This was a multi-classification task, i.e., to classify the emotions into seven labels. Here also, we fine-tuned RoBERTa model (Liu et al., 2019) with the same five layers, just changing the output neurons to 7 instead of 1. We had used Cross-Entropy Loss as the loss function (which already has a softmax layer). We also upsampled the dataset as it was imbalanced. Highly imbalanced data poses added difficulty, as most learners will exhibit bias towards the majority class and, in extreme cases, may ignore the minority class altogether (Johnson and Khoshgoftaar, 2019). Random over-sampling (Moreo et al., 2016) was performed using the imblearn library. The imbalanced dataset can be seen in figure 2, the minority class being the emotion labeled "joy".

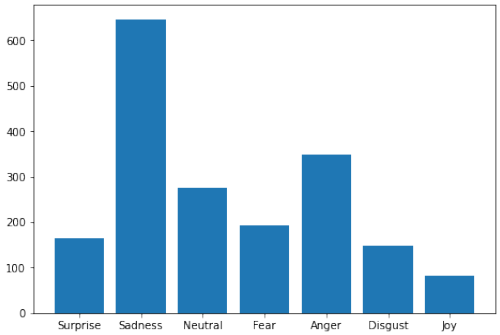

Figure 2: Imbalanced Dataset

4.3 Hyperparameters and other settings

For all the tasks, the learning rate was set to 10^{-5}, and the models were trained using Adam (Kingma and Ba, 2014) as optimizer. The parameters of Adam were Beta(0.9, 0.999) and weight decay as 0. The batch size was set to 8. The dataset was shuffled using Pytorch (Paszke et al., 2019) data loader. All the models were trained on the GPUs provided by Google Colab.

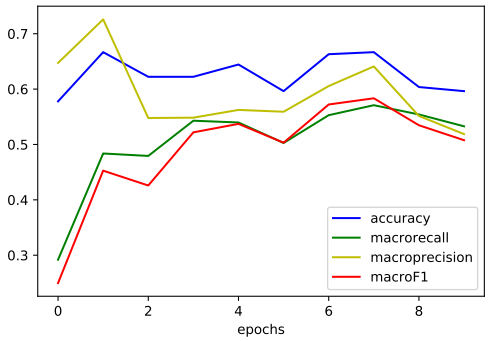

Figure 3: Metrics used for EMO task

5 Results

Our system achieved a Pearson Correlation Coefficient (Benesty et al., 2009) of 0.488 on Task 1. Empathy Pearson Correlation was 0.470, and Distress Pearson Correlation was 0.506. Hence, the average of both was taken as the final score. In the development set, the empathy score was 0.4583 (after the 8th epoch), and the distress score was 0.4415 (after the 4th epoch, as after the score was decreased due to overfitting). Although the empathy score was slightly high, it yielded less score in the test set due to overfitting. While due to early stopping, distress yielded a better score.

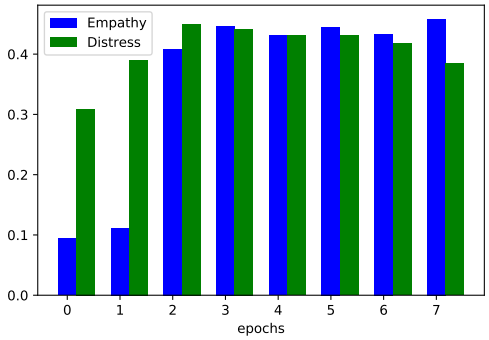

Figure 4: Empathy and Distress scores (Augmented)

We had two different submissions for the emotion classification, one with augmentation and upsampling and one without altering the data. The test scores of both submissions are mentioned in

Table 2. Also, the results of the development set are plotted in figure 4. We tested until ten epochs but decided to submit the model, trained only up to eight epochs as it was overfitting. Hence, the macro F1-score decreased on the development set despite accuracy increasing on the training set.

6 Conclusion

This paper describes our submission to the WASSA 2022 shared task, where we have used the already trained RoBERTa model on a large dataset and then used its power by just finetuning on the given dataset. By the approach we have used, it can also be deduced that text augmentation and upsampling helped in emotion classification and predicting the empathy and distress scores as most of the time, the larger amount of data helps improve the training process of a model.

References

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Sarah Fabi, Lydia Anna Weber, and Hartmut Leuthold. 2019. Empathic concern and personal distress depend on situational but not dispositional factors. *PloS one*, 14(11):e0225102.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A Smith. 2021. Finetuning pretrained transformers into rnns. *arXiv preprint arXiv:2103.13076*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Meera Komarraju, Steven J Karau, Ronald R Schmeck, and Alen Avdic. 2011. The big five personality traits, learning styles, and academic achievement. *Personality and individual differences*, 51(4):472–477.

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. *arXiv preprint arXiv:1907.11692*.

Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. 2016. Distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 805–808.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jonathan J Webster and Chunyu Kit. 1992. Tokenization as the initial phase in nlp. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.