SURREY-CTS-NLP at WASSA2022: An Experiment of Discourse and Sentiment Analysis for the Prediction of Empathy, Distress and Emotion

Shenbin Qian[1], **Constantin Orasan**[1], **Diptesh Kanojia**[2],
Hadeel Saadany[1] and Felix do Carmo[1]
Centre for Translation Studies[1],
Department of Computer Science[2],
University of Surrey, UK
{s.qian, c.orasan, d.kanojia, h.saadany, f.docarmo}@surrey.ac.uk

Abstract

This paper summarises the submissions our team, SURREY-CTS-NLP has made for the WASSA 2022 Shared Task for the prediction of empathy, distress and emotion. In this work, we tested different learning strategies, like ensemble learning and multi-task learning, as well as several large language models, but our primary focus was on analysing and extracting emotion-intensive features from both the essays in the training data and the news articles, to better predict empathy and distress scores from the perspective of discourse and sentiment analysis. We propose several text feature extraction schemes to compensate the small size of training examples for fine-tuning pre-trained language models, including methods based on Rhetorical Structure Theory (RST) parsing, cosine similarity and sentiment score. Our best submissions achieve an average Pearson correlation score of 0.518 for the empathy prediction task and an F1 score of 0.571 for the emotion prediction task[1], indicating that using these schemes to extract emotion-intensive information can help improve model performance.

1 Introduction

Large transformer models (Vaswani et al., 2017) have shown their power in various natural language processing (NLP) downstream tasks, especially in dealing with informative text. However, for text containing human emotions, current models still need to be improved and trained on more emotion-intensive datasets. Empathy and emotion prediction has gained a lot of attention in the field of NLP with many shared tasks and challenges being proposed in recent years.

For the WASSA 2022 Shared Task, we have participated in two of their 4 tracks, which are:

Track 1: Empathy Prediction (EMP), which is a regression task to predict both the empathy and distress score at the essay-level.
Track 2: Emotion Classification (EMO), which is to classify each essay into one of seven classes of emotion.

Both tracks are supposed to use the same dataset the organisers provide, which we will discuss in the next section. In Section 2, we explore some interesting features of the dataset and show what methods and strategies we have paid closer attention to, according to the data features. Section 3 gives a detailed introduction to the schemes we use, as well as different learning strategies we adopt for analysing the dataset and for incorporating additional features to train our models. Section 4 shows results of our proposed methods, as well as future directions that would be interesting to explore. In Section 5, we present our conclusions and summarise our methods.

2 Initial Data Analysis

The original data used in this shared task were gathered for experiments to predict empathy based on Batson's Empathic Concern and Personal Distress Scale (Batson et al., 1987). Participants were given news articles to read and then wrote a short essay to describe how they feel about the news. Thereafter, they were given questions to answer, which were designed for grading their empathy and distress from level 1 to 7. The demographic and personality information of these participants were also collected for further studies on how these factors might affect their empathy and distress level. The emotion labels which annotate the data were produced semi-automatically: human annotators corrected the automatic predictions of deep learning models. More details of how this dataset was designed can be found in (Buechel et al., 2018) and

[1]The organisers have not yet released the official result and ranking on the leaderboard when this paper is written.

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 271 - 275

(Tafreshi et al., 2021).

After a quick exploration of the dataset, we noticed that the training size is very small, compared to the size of the datasets used in modern transformer models, with only 2130 examples in total, including the development dataset. Due to the designing purpose of the empathy prediction task, the majority of these selected news articles are negative in nature so as to induce the annotators' empathy. However, this leads to a skewed distribution for emotion classification (see Figure 1), which might influence the prediction of the minority classes.

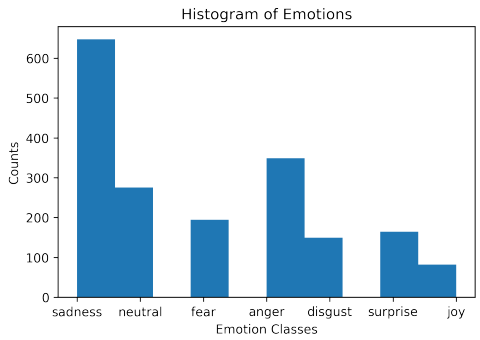

Figure 1: Distribution of Emotion Classes

Another feature in the dataset which could act as a good predictor of empathy and distress is demographic and personality information, since people from various backgrounds and with different personalities may have different views and feelings towards these news articles. We found that some variables like personality agreeableness do have a relative correlation with the empathy score (see Table 1). Therefore, we opted for incorporating this information with text as additional features.

Personality Extraversion	Personality Agreeableness
0.209025617	0.243257229

Table 1: Pearson Correlation between Empathy and Some Personality Information

From both Batson's Empathy Theory and the high Pearson correlation score (0.45) of empathy and distress, we know that the two variables are highly correlated. Therefore, multi-task learning could help us learn features from the empathy prediction task to apply to the distress prediction task.

The most important thing we learnt from this dataset, which can help supplement the lack of adequate training data, is that the essays are the responses to the news articles. We, therefore, put forward the assumption that the news article must contain features that trigger the emotion of the reader. We can regard the news article and the essay as one unified discourse, where some parts are more emotion-intensive, while others are more descriptive than emotional. Thus, we explored methods adopted for both discourse and sentiment analysis to extract emotion-intensive features from the articles to help with the prediction.

3 Methods Description

3.1 Empathy Prediction

We tried different approaches to extract features that indicate emotions from the text, namely, RST (Mann and Thompson, 1987) parsing, cosine similarity and sentiment score. We also included demographic and personality information to train a tabular transformer model to see if this information would help the prediction. Multi-task learning was also used to train one model for both the empathy and distress sub-tasks.

3.1.1 RST Parsing

Rhetorical structure theory aims to build a tree which represents the discourse structure for a sequence of text units. In such a tree structure, we know that units defined as nuclei of a rhetorical relation are more essential to the writer's purpose, while those defined as satellites would become incomprehensible if nuclei were deleted (Mann and Thompson, 1987). In our case, we assumed that in the essays there are some parts that are more emotional, carrying the intention of the writer, i.e. the annotator, whereas others are only a rephrasing of the events in the corresponding news article in a descriptive way. We also made a further assumption that nuclei should be given more weights on the text embeddings while satellites less weights during the training process.

In the experiments, we used the text-level discourse rhetorical structure (DRS) parser by Zhang et al. (2021), which uses adversarial learning to generate DRS trees from a top-down global perspective, and claims to be one the state-of-the-art parsers in this area. We gave different weights to the embeddings of nuclei and satellites and found that giving 0.3 to the nuclei and 0.7 to the entire essays for fine-tuning a RoBERTa base model (Liu et al., 2019) leads to our best performance. In the experiments, we used an AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 0.00002.

3.1.2 Cosine Similarity and Sentiment Score

Since these news articles are long and some of them are mixed with URLs and other noise like missing content[2], using an RST parser to get their discourse tree is not likely to produce useful information and hence not a feasible approach for feature extraction. For this reason, our goal was to extract those sentences that are highly related to the essays from the articles.

Sentence embeddings represent sentences as numerical vectors which represent the semantic information of the sentence. For this reason cosine similarities between sentence embeddings of the essays and the articles can be calculated to extract sentences in the articles that are semantically similar to those of the essays (see Equation 1, where u is the sentence embeddings for the article and v for the essay). Also, sentiment scores were used to extract sentences in the articles that contain more extreme sentiments.

$$Cosine\_similarity = 1 - \frac{u \cdot v}{||u||_2||v||_2} \quad (1)$$

To get cosine similarities between sentences, we tried two sentence-level embedders, e.g. Sentence-BERT (Reimers and Gurevych, 2019) and Universal Sentence Encoder (Cer et al., 2018). The latter was used in our final model. For the calculation of sentiment scores, we used a simple rule-based sentiment analysis tool, VADER (Hutto and Gilbert, 2014), which claims to achieve 0.96 in F1 score for sentiment classification. Cosine similarity and sentiment score can be used together or separately to extract features in the articles. We experimented different thresholds to filter sentences in the articles and concatenate them with essays. In our final model, sentences with cosine similarity higher than 0.2 and sentiment score higher than 0.6 or lower than -0.6 are kept, so that a reasonable amount of sentences which are semantically similar to the essays and sentimentally extreme can be fed into our model.

3.1.3 Tabular Models and Multi-task Learning

Demographic and personality information were used together with essays and articles to train a tabular model based on Gu and Budhkar (2021), and we got the highest Pearson correlation score (0.53) in empathy prediction during training. However, as personality information is not included in the test data, we are not able to submit the result of this approach to the Shared Task.

A weighted loss considering the homoscedastic uncertainty (Kendall et al., 2017) of our two sub-tasks was applied to our RoBERTa model (Liu et al., 2019) to predict both empathy and distress for multi-task learning. We used the same hyperparameters as in the model of RST parsing, but trained it with more epochs to minimise their shared loss.

3.2 Emotion Prediction

For the emotion classification task, we also tested those methods in empathy prediction, but the results are not as good as expected during our training process. Therefore, we adopted data augmentation and ensemble learning to improve model performance.

3.2.1 Data Augmentation with GoEmotions Dataset

As the original training data is small in size and relatively skewed in distribution, data augmentation is something that we could do to overcome the problems. The GoEmotions dataset (Demszky et al., 2020) is a manually annotated high-quality dataset with 27 emotion categories based on 58k English Reddit comments, making itself a good source for data augmentation. However, as texts in the GoEmotions dataset might have different writing styles and sequence lengths compared with our essays, we cannot simply use all the data to train our model. We selected those texts that are longer than 25 words and make sure that more joy and surprise examples are included to compensate the skewed distribution.

3.2.2 Ensemble Learning

Trying larger models or combining the results of several different models would be another way to compensate the small training size. Ensemble learning is a machine learning strategy that combines the prediction of multiple algorithms to get better performance. For this task, we fine-tuned the RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020) and DeBERTa (He et al., 2020) base models for majority voting to get a better predictive result.

[2]We list some of these problems in Appendix A.

	RST Parsing	Similarity & Sentiment Score	Multi-task Learning	Simple Fine-tuning
Empathy	0.431	0.501[3]	0.480	0.504
Distress	0.465	0.535	0.458	0.530

Table 2: Pearson Correlation of Predicted Empathy and Distress Scores

4 Results and Discussions

4.1 Results for Empathy Prediction

Table 2 compares the results based on RST parsing, cosine similarity and sentiment score, multi-task learning and simple fine-tuning. The Pearson correlation is calculated using the evaluation script provided by the organisers on the test dataset.

We can see that the Pearson correlation scores produced by the model using RST parser are not as high as expected, but results using extracted article sentences by cosine similarity and sentiment score are pretty high, especially the distress score. However, just fine-tuning a RoBERTa base model also achieves high scores. This indicates that there do exist features in the article that trigger the feeling of the reader but we need to better analyse and extract these features from the articles. Multi-task learning is also not bad at predicting the empathy score, but we might still need to design a better loss function to train the model.

For future directions, RST parsing or even other methods for discourse analysis is still something we can try to get useful information from the articles.

4.2 Results for Emotion Prediction

Table 3 lists the result of using the GoEmotions dataset as additional training data, the result for ensemble learning mentioned in Section 3.2, as well as the result of simply fine-tuning a RoBERTa base model.

	GoEmotions	Ensemble Learning	Simple Fine-tuning
Accuracy	0.634	0.619	0.646
F1 score	0.548	0.534	0.571
Precision	0.576	0.564	0.595
Recall	0.532	0.520	0.559

Table 3: Scores for Emotion Prediction

We see that the F1 score for the GoEmotions result is higher than the one for ensemble learning, which implicitly suggests that getting more training data is more important than using larger and more models, especially when training datasets are particularly small. However, just fine-tuning a RoBERTa base model appears to have a slightly better result than data augmentation in this task. This could be related to how we sample the dataset, since data augmentation might make the training data have a very different distribution from the test data.

For future directions, how to get and sample extra data to compensate the skewed distribution or experimenting with feature extraction techniques on existing information in the training data like the news articles or demographic information could be possible ways to improve model performance.

5 Conclusions

This paper summarises the submissions our team has made to the WASSA 2022 Shared Task for empathy, distress and emotion prediction. In this work, we tried different ways to improve model performance from the perspective of discourse and sentiment analysis, data augmentation and method optimisation like RST parsing, sentiment score and ensemble learning. We propose a reliable method to analyse and extract information from both the news articles and the essays to compensate the small training size for empathy and distress prediction, that is, using similarity and sentiment scores for feature extraction. Adding GoEmotions (Demszky et al., 2020) data to increase the training size is one way to improve emotion prediction, but attention should be paid to how much data we should sample for each category. In our best submission, we get a Pearson correlation score of 0.518 for the empathy prediction task and an F1 score of 0.571 for the emotion prediction task.

The method we used to extract emotion-intensive features is by no means perfect, future studies could explore other methods in discourse or text analysis to further improve model performance when dealing with emotion data with a small training size.

[3]Only this result is based on fine-tuning a RoBERTa large model, not the base model

References

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively dis-

article_id	problem	response_id	empathy	distress	emotion
63	missing content	R_1DAmmWVuxekOzQt	4	1	surprise
36	one sentence news	R_3oZwv1aOvzgfBPT	5.5	1	sadness
142	two different articles as one	R_1rfDsNtkx9ueNuH	1	1	anger
412	mixed with URL				

Table 4: Problems of News Articles

tinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. pages 4758–4765.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *arXiv preprint*.

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

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

Ken Gu and Akshay Budhkar. 2021. Multimodal-toolkit: A package for learning on tabular and text data with transformers. pages 69–73. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint*.

C J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–225.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint*.

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*.

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

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, João Sedoc, Sven Buechel, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. pages 92–104. Association for Computational Linguistics.

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint*.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. pages 3946–3957. Association for Computational Linguistics.

A Appendix

We randomly read some of the news articles and find several problems that might affect participants' responses and thus undermine their empathy and emotion. We list these problems in Table 4.