# Curtin OCAI at WASSA 2023 Empathy, Emotion and Personality Shared Task: Demographic-Aware Prediction Using Multiple Transformers

**Md Rakibul Hasan**[1] **Md Zakir Hossain**[1] **Tom Gedeon**[1]
**Susannah Soon**[1] **Shafin Rahman**[2]
[1]Optus Centre for AI, Curtin University, Perth WA 6102, Australia
[2]North South University, Dhaka 1229, Bangladesh
{rakibul.hasan,zakir.hossain1,tom.gedeon,susannah.soon}@curtin.edu.au
shafin.rahman@northsouth.edu

## Abstract

The WASSA 2023 shared task on predicting empathy, emotion and other personality traits consists of essays, conversations and articles in textual form and participants' demographic information in numerical form. To address the tasks, our contributions include (1) converting numerical information into meaningful text information using appropriate templates, (2) summarising lengthy articles, and (3) augmenting training data by paraphrasing. To achieve these contributions, we leveraged two separate T5-based pre-trained transformers. We then fine-tuned pre-trained BERT, DistilBERT and ALBERT for predicting empathy and personality traits. We used the Optuna hyperparameter optimisation framework to fine-tune learning rates, batch sizes and weight initialisation. Our proposed system achieved its highest performance – a Pearson correlation coefficient of 0.750 – on the conversation-level empathy prediction task[1]. The system implementation is publicly available at https://github.com/hasan-rakibul/WASSA23-empathy-emotion.

## 1 Introduction

Empathy refers to an individual's capacity to comprehend and express appropriate emotions in response to others' emotions, perspectives and beliefs (Decety and Jackson, 2004). This ability can foster relationships and reduce stress and unhappiness among individuals through interaction. The importance of empathy is evident across a broad range of real-life human interactions, such as patient-doctor (Jani et al., 2012), teacher-student (Aldrup et al., 2022) and human-robot (Spitale et al., 2022) interactions.

The Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) has organised a "Shared Task on Empathy Detection, Emotion Classification and Personality Detection in Interactions" in 2023 (Barriere et al., 2023). The challenge involves predicting empathy, emotion and personality traits from two types of datasets: essay and conversation. The essay-level dataset consists of essays written by study participants in response to news articles involving harm to individuals, groups or other entities. The conversation-level dataset includes textual conversations between participants regarding the news articles. In addition to the textual data (essays and conversations), the datasets also provide demographic and personal information in numerical form. We participated in four tracks of the 2023 challenge, which involves predicting (1) empathy, personality and interpersonal reactivity index from the essay-level dataset and (2) empathy and emotion from the conversation-level dataset.

WASSA 2023 challenge extends from the 2022 challenge (Barriere et al., 2022) that involved predictions from only an essay-level dataset. Participants in 2022 challenge, such as Vasava et al. (2022); Chen et al. (2022); Qian et al. (2022); Del Arco et al. (2022); Lahnala et al. (2022) and Ghosh et al. (2022), employed transformer-based architectures, such as BERT (Devlin et al., 2018). Transformer-based models were also found to be the best-performing model in the WASSA 2021 shared task on empathy prediction (Tafreshi et al., 2021). Apart from WASSA competition, transformer models are also used in predicting empathy in essays written by medical students about simulated patient-doctor interactions (Dey and Girju, 2022).

Transformer models are deemed highly suitable for undertaking text-based empathy prediction owing to their inherent ability to effectively capture long-range dependencies through attention mechanism (Vaswani et al., 2017). Fine-tuning pre-trained transformers harnesses prior knowledge,

[1]At the time of writing this paper, official rankings on any tasks and evaluations of several tasks in which we participated have not been published yet.



*Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 536–541

leading to enhanced performance while minimising training time. Qian et al. (2022) reported the best performance by just fine-tuning a BERT-based model in their system for the WASSA 2022 shared task. We, therefore, choose to fine-tune pre-trained transformers to predict empathy and personality traits in this challenge. In our prediction pipeline, we utilise numerical information from the datasets, such as participants' demographic information and income, because previous research by Guda et al. (2021) showed demographic information is an important cue in text-based empathy prediction.

Overall, this paper has made the following contributions: (1) we use novel strategies to incorporate numerical demographic and other data in the text-based prediction pipeline, (2) we summarise longer text sequences to fit into the pipeline, and (3) we augment training samples by paraphrasing the textual data.

## 2 System description

The general prediction system for essay-level tasks is illustrated in Figure 1. In the case of conversation-level tasks, demographic and other personal information are not available in the conversation-level dataset. In that case, our prediction models involve only conversations and summarised articles, followed by paraphrasing to augment the training dataset.

### 2.1 Number to text mapping

We first discarded data points from the datasets where any component is missing. The data collection process, along with the questionnaires used in the WASSA 2023 datasets, has been detailed in the work of Omitaomu et al. (2022). Based on the reported distribution of demographic information, we have mapped numerical values of gender, education level and race to their corresponding textual information as illustrated in Table 1.

All the textual features were concatenated in the order of appearance, and this combined feature is referred to as the *demographic* feature throughout this paper. We further concatenated the *demographic* feature with the *essay* texts to create the *demographic_essay* feature.

### 2.2 Article summarisation

The converted article text comprised long sequences with a maximum length of 20,047 characters. In contrast, the *demographic_essay* feature had a maximum of 956 characters, resulting in 236 tokens. Since the BERT tokeniser we used can process a maximum of 512 tokens, the entire article text cannot be processed in its current form. Consequently, we generated summaries of the articles. We employed *flan-t5-base-samsum*[2], which is a fine-tuned variant of the model proposed by Chung et al. (2022).

The maximum length of the summarised articles was 987 characters. Considering that the *demographic_essay* feature contained 956 characters, resulting in 236 tokens, it seems plausible that incorporating the additional 987 characters of the article summary would be within the limit of BERT's maximum token length of 512.

| Numeric feature | Converted text |
|---|---|
| gender | I am <gender>. |
| age | My age is <age> years. |
| education level | My education level is <education level>. |
| race | My race is <race>. |
| income | My income is <income>. |
| article_id | I read newspaper article <article_id>. |

Table 1: Templates used to transform numerical features into meaningful texts.

### 2.3 Data augmentation

In order to augment the number of training samples, we utilised the *chatgpt_paraphraser_on_T5_base*[3] to paraphrase the *demographic*, *essay* and *article* texts, effectively doubling the size of the dataset.

### 2.4 Model and hyperparameter tuning

We experimented with different hyperparameter configurations illustrated in Table 2. Specifically, we fine-tuned three transformer models from Huggingface (Wolf et al., 2019). In fine-tuning BERT-based models, weight initialisation plays a critical role (Dodge et al., 2020). Therefore, we also explored various seed values for CPU and GPU. For conversation-level tasks, the length of the conversation texts was comparatively shorter than that of essay-level tasks. Consequently, we investigated larger batch sizes in the range of 2 to 16.

[2] https://huggingface.co/philschmid/flan-t5-base-samsum

[3] https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

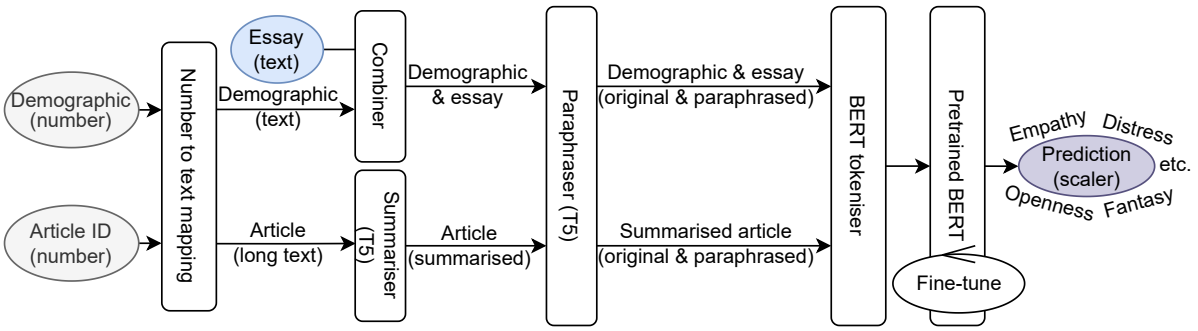



Figure 1: Overall system for essay-level tasks. First, we map numerical features into meaningful text. Next, we leverage a T5-based model to summarise lengthy articles. We use paraphrasing as a data augmentation technique. Finally, we fine-tune a pre-trained BERT model to predict the degree of empathy and other personality traits.

| Hyperparameter | Search space |
|---|---|
| Model | bert-base-uncased, distilbert-base-uncased, albert-base-v2 |
| Learning rate | $[10^{-05} - 10^{-04}]$ |
| Batch size | $[2 - 8]$ |
| Seed | $[1 - 100]$ |

Table 2: Hyperparameter tuning search space for essay-level tasks.

To tune the hyperparameters, we utilised Optuna (Akiba et al., 2019), with the default tree-structured Parzen estimator as the sampler and the median stopping rule as the pruner. The purpose of the pruner is to stop the tuning process on low-performing hyperparameters early, both to save resources and to enable a greater focus on the best-performing hyperparameters.

The best model, as determined by Optuna, was fine-tuned separately for each of the 14 regression tasks we participated. We employed the Pytorch AdamW optimiser with a default weight decay of 0.01 and betas of 0.9 and 0.999 to optimise the mean-squared-error loss function. To adjust the learning rate, we utilised a linear learning rate scheduler with zero warmup steps. We evaluated the prediction performance of all regression tasks in terms of the official Pearson correlation coefficient metric. We trained all essay-level models for 35 epochs and conversation-level models for 50 epochs. We determined the optimal number of epochs by monitoring the training loss until convergence was reached. We observed that the conversation-level dataset required more epochs for convergence, likely due to its larger size compared to the essay-level dataset.

### 2.5 Resources

We trained the model on a Tesla V100 32 GB GPU and used the following software packages: Transformers 4.28.1, Datasets 2.12.0, Pytorch 2.0.0, CUDA 11.8, Optuna 3.1.1, Numpy 1.24.3, Pandas 1.5.3, Plotly 5.14.1 with Python 3.10.10.

## 3 Result & analysis

To determine which feature sets are most effective for predicting empathy, we conducted an experiment in which we combined different features (*essay*, *demographic*, *demographic_essay* and *article*) and trained a DistilBERT (Sanh et al., 2019) model using 5-fold cross-validation for 10 epochs. Huggingface's tokeniser allowed us to tokenise pairs of sequences together by automatically concatenating them with a special [SEP] token. We then used these pairs of features and evaluated their performance, as presented in Table 3.

| Features | Average Pearson $r$ |
|---|---|
| demographic_essay-article (long) | 0.819 |
| demographic_essay-article (summary) | 0.865 |
| essay-demographic | 0.807 |
| essay-article | 0.565 |

Table 3: Five-fold cross-validated (combined training and development set) essay-level empathy prediction performance using various input features in a DistilBERT model. The *demographic_essay* feature refers to manually concatenated pairs of *demographic* and *essay* texts, while the hyphenated features, such as *essay-demographic*, denote automated concatenation by the tokeniser.

Conversion from the longer version of the *article* text to its summarised shorter version improved the performance (Table 3). We speculate that the reason for comparatively lower performance

with longer articles is the BERT tokeniser's limitation in accommodating longer texts. The inclusion of *demographic* and *article* features with the *essay* feature improved the model's overall performance. Therefore, we have incorporated *demographic_essay* and *article* features in our final model for the essay-level tasks.

It is worthwhile to note that the use of data augmentation techniques such as paraphrasing can introduce very similar samples in the dataset. It may bias the evaluation metrics, especially when similar samples are present in both the training and validation sets. The cross-validated Pearson correlation coefficient reported in Table 3 includes both the training and development sets with data augmentation (paraphrasing). However, in the process of tuning the model hyperparameters, we only used paraphrasing with the training set and not with the development set to prevent any potential bias caused by the duplication of similar samples.

Among the pre-trained transformer models we experimented with (BERT, DistilBERT and ALBERT), the BERT base model was the best-performing model. Accordingly, we used BERT and tuned the other hyperparameters. We conducted 200 and 100 Optuna trials for essay-based empathy and distress prediction models, respectively. As the best set of hyperparameters is always found within the first 50 trials in the essay-level empathy and distress prediction models, 50 trials were run for other essay-level prediction models. In the case of conversation-level tasks, 100 Optuna trails were run. Table 4 presents the best set of hyperparameters found by the Optuna trials.

| Prediction task | Learning rate | Batch size | Seed | Pearson $r$ |
|---|---|---|---|---|
| Empathy | 4.27e-05 | 5 | 1 | 0.785 |
| Distress | 1.85e-05 | 6 | 6 | 0.726 |
| Conscientiousness | 5.98e-05 | 7 | 30 | 0.791 |
| Openness | 1.80e-05 | 2 | 81 | 0.776 |
| Extraversion | 1.61e-05 | 7 | 34 | 0.681 |
| Agreeableness | 5.15e-05 | 6 | 65 | 0.819 |
| Stability | 5.36e-05 | 5 | 13 | 0.627 |
| Perspective taking | 4.30e-05 | 2 | 65 | 0.837 |
| Personal distress | 4.38e-05 | 7 | 56 | 0.788 |
| Fantasy | 5.36e-05 | 5 | 13 | 0.895 |
| Empathic concern | 4.92e-05 | 5 | 1 | 0.850 |
| Emotional polarity[a] | 1.06e-05 | 10 | 96 | 0.763 |
| Emotion[a] | 1.44e-05 | 10 | 87 | 0.768 |
| Empathy[a] | 1.97e-05 | 12 | 68 | 0.711 |

[a]Conversation-level

Table 4: Optimal hyperparameters tuned by Optuna and their evaluation (Pearson correlation coefficient) on the original development set without overlapping samples due to augmentation.

We investigated the relative importance of learning rate, seed and batch size (see Appendix A). Our findings are consistent with prior research by Dodge et al. (2020), which highlighted the impact of seed value on the fine-tuning performance of BERT-based models. However, the relative importance of hyperparameters varied across the prediction tasks, indicating the task-specific nature of fine-tuning pre-trained transformer models. It guided us to train separate models for separate tasks.

We observed that text summarisation and data augmentation (paraphrasing) improved model performance on the development set. On the test dataset, the final model achieved Pearson correlation coefficients of 0.750, 0.683 and 0.573 for conversation-level empathy, emotional polarity and emotional intensity prediction, respectively. For essay-level tasks, we achieved Pearson correlation coefficients of 0.187 and 0.344 for empathy and distress predictions, respectively. The average Pearson correlation coefficient for conversation-level tasks was 0.669, while it was 0.266 for essay-level empathy and distress prediction. The test performance of essay-level personality and interpersonal reactivity index predictions, as well as the official rankings, have not been published at the time of writing this paper. Nevertheless, our system achieved its best performance of a Pearson correlation coefficient of 0.750 in predicting conversation-level empathy.

## 4 Conclusion

Empathy is a vital human attribute to support and care for others. This paper outlines a comprehensive system for predicting empathy, emotion and other personality traits as part of the WASSA 2023 shared task. To this end, we first map the numerical demographic information into meaningful text since individuals' demographic information, such as age, sex and race, may affect their empathic capacity. Our system utilises pre-trained transformers to map numerical information into meaningful text, summarise longer text sequences, paraphrase text sequences to augment smaller training datasets and finally predict the degree of empathy and other personality traits.

## Acknowledgements

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS-enabled capability supported by the Australian Government.

## References


Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Karen Aldrup, Bastian Carstensen, and Uta Klusmann. 2022. Is empathy the key to effective teaching? a systematic review of its association with teacher-student interactions and student outcomes. *Educational Psychology Review*, 34(3):1177–1216.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. WASSA 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. IUCL at WASSA 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 228–232, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Flor Miriam Del Arco, Jaime Collado-Montañez, L. Alfonso Ureña, and María-Teresa Martín-Valdivia. 2022. Empathy and distress prediction using transformer multi-output regression and emotion analysis with an ensemble of supervised and zero-shot learning models. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Priyanka Dey and Roxana Girju. 2022. Enriching deep learning with frame semantics for empathy classification in medical narrative essays. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 207–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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*.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team IITP-AINLPML at WASSA 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 255–260, Dublin, Ireland. Association for Computational Linguistics.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Bhautesh Dinesh Jani, David N Blane, and Stewart W Mercer. 2012. The role of empathy in therapy and the physician-patient relationship. *Complementary Medicine Research*, 19(5):252–257.

Allison Lahnala, Charles Welch, and Lucie Flek. 2022. CAISA at WASSA 2022: Adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 280–285, Dublin, Ireland. Association for Computational Linguistics.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Shenbin Qian, Constantin Orasan, Diptesh Kanojia, Hadeel Saadany, and Félix Do Carmo. 2022. SURREY-CTS-NLP at WASSA2022: An experiment of discourse and sentiment analysis for the prediction of empathy, distress and emotion. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 271–275, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Micol Spitale, Sarah Okamoto, Mahima Gupta, HAO Xi, and Maja J Matarić. 2022. Socially assistive robots as storytellers that elicit empathy. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4):1–29.

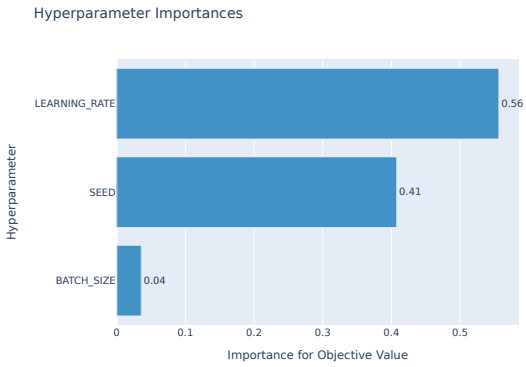
Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264, Dublin, Ireland. Association for Computational Linguistics.

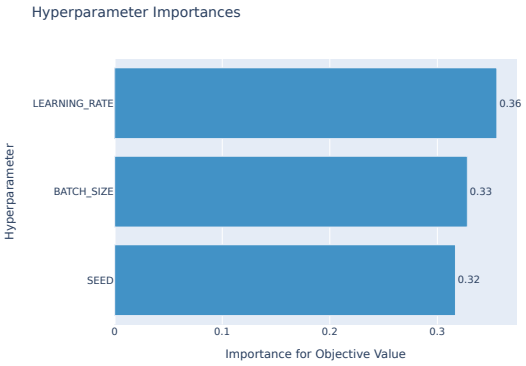
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

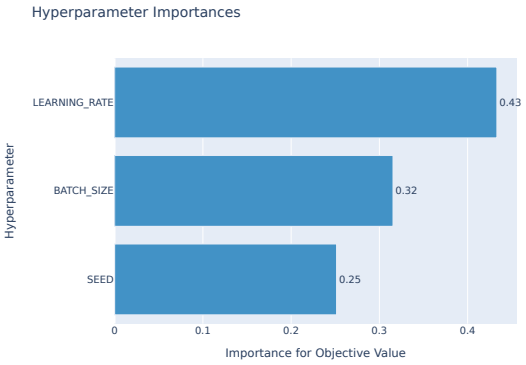
## Appendix A Hyperparameter importance



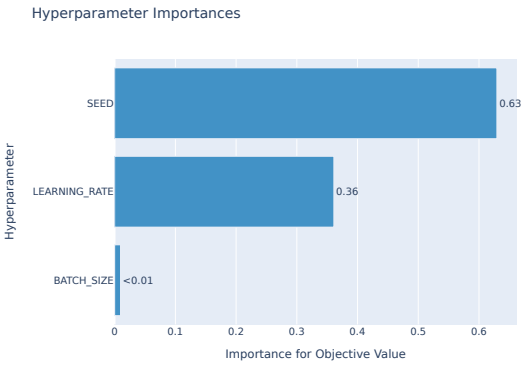

(a) Empathy prediction: learning rate has the highest impact (56% relative importance), followed by the seed value. Batch size is the least impactful (4% relative importance).




(b) Conscientiousness prediction: learning rate, batch size and seed value all have a high impact with a relative importance of around 30%.




(c) Personality distress prediction: learning rate has the highest impact (43% relative importance), followed by batch size (32% relative importance) and seed value (25% relative importance).




(d) Empathic concern prediction: Seed value has the highest impact, followed by the learning rate. Batch size is the least impactful, having less than 1% relative importance.

Figure A1: Relative importance of learning rate, seed value and batch size in various essay-level tasks. Here, the objective value refers to the Pearson correlation coefficient. The variations in hyperparameter importance across tasks indicate the requirements of training separate models for separate tasks.