

UPLME: Uncertainty-Aware Probabilistic Language Modelling for Robust Empathy Regression

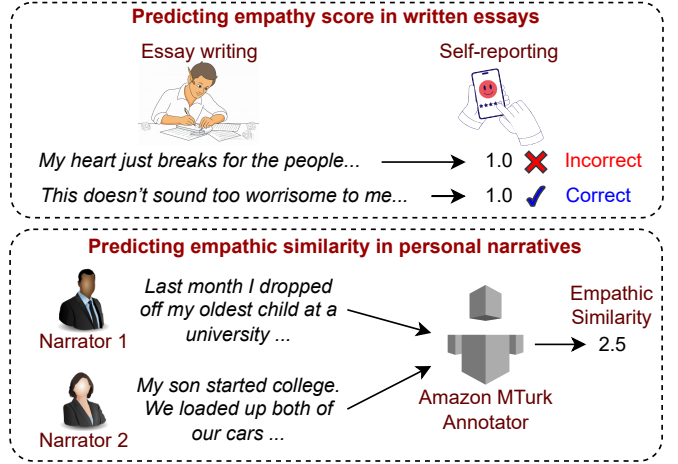
Md Rakibul Hasan, *Graduate Student Member, IEEE*, Md Zakir Hossain, *Member, IEEE*, Aneesh Krishna, Shafin Rahman, and Tom Gedeon, *Senior Member, IEEE*

Abstract—Noisy self-reported empathy scores challenge supervised learning for empathy regression. While many algorithms have been proposed for learning with noisy labels in textual *classification* problems, the *regression* counterpart is relatively under-explored. We propose *UPLME*, an uncertainty-aware probabilistic language modelling framework to capture label noise in empathy regression tasks. One of the novelties in UPLME is a probabilistic language model that predicts both empathy scores and heteroscedastic uncertainty, and is trained using Bayesian concepts with variational model ensembling. We further introduce two novel loss components: one penalises degenerate Uncertainty Quantification (UQ), and another enforces similarity between the input pairs on which empathy is being predicted. UPLME achieves state-of-the-art performance (Pearson Correlation Coefficient: $0.558 \rightarrow 0.580$ and $0.629 \rightarrow 0.634$) in terms of the performance reported in the literature on two public benchmarks with label noise. Through synthetic label noise injection, we demonstrate that UPLME is effective in distinguishing between noisy and clean samples based on the predicted uncertainty. UPLME further outperform (Calibration error: $0.571 \rightarrow 0.376$) a recent variational model ensembling-based UQ method designed for regression problems. Code is publicly available at <https://github.com/hasan-rakibul/UPLME>.

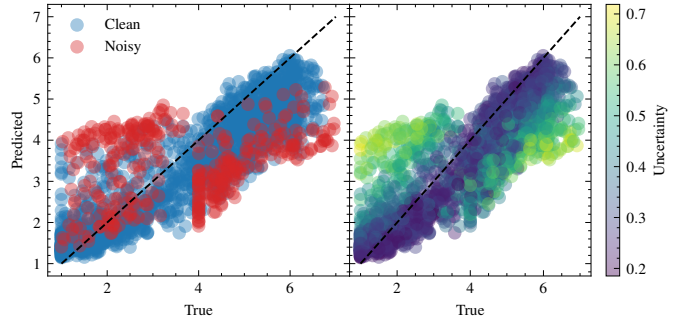
Index Terms—Uncertainty quantification, heteroscedastic, empathy regression, noisy label, robust learning.

I. INTRODUCTION

Empathy is an involuntary, vicarious reaction to another person’s feelings or circumstances [1], which drives supportive communication across various human-human [2] and human-machine [3] interactions. Automatic empathy regression from textual content has gained traction over the past few years [2]. In most empathy regression studies, the “*regression*” part refers to estimating the response (i.e., an aggregated continuous value) of empathy measurement scales or questionnaires that would otherwise be used for a manual assessment of empathy. Over many years, social psychology literature has developed various such empathy measurement scales. One such scale is developed by Batson et al. [4], which measures empathy in terms of six dimensions: sympathetic, moved,



(a) **Top**: predicting empathy scores in written essays and an example of how label noise arises in this task: self-reported scores (an empathy score of 1.0 means the lowest empathy) did not align with the corresponding essay. **Bottom**: predicting empathic similarity, where our approach is also effective.



(b) **Left**: label noise injected (labels are shifted by 3.0) in 30% samples, most of which are mispredicted by the model. **Right**: predicted uncertainty is higher on incorrect predictions, and uncertainties correlate with noise.

Fig. 1. (a) Demonstration of empathy regression tasks and (b) Uncertainty based on our proposed penalty loss capturing label noise by ensuring higher uncertainty estimation on noisy samples and lower on clean samples.

compassionate, tender, warm and soft-hearted. Being a subjective phenomenon and further to conform to the definitions of empathy developed in social psychology literature, training labels for automated empathy detection systems often come from self-assessed empathy labels [2]. The de facto method of creating such datasets for empathy detection is crowdsourcing, which often results in incorrect empathy labels due to the nature of crowdsourcing [5, 6]. While circumventing this issue of label noise, building a robust empathy detection system is a long-standing challenge.

M R Hasan, M Z Hossain, A Krishna and T Gedeon are with School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley WA 6102, Australia.

S Rahman is with North South University, Dhaka 1229, Bangladesh.

M R Hasan is also with BRAC University, Dhaka 1212, Bangladesh.

M Z Hossain and T Gedeon are also with The Australian National University, Canberra ACT 2600, Australia.

T Gedeon is also with University of Óbuda, 1034 Budapest, Hungary.

E-mail: {rakibul.hasan, zakir.hossain1, a.krishna, tom.gedeon}@curtin.edu.au, shafin.rahman@northsouth.edu

Corresponding author: M R Hasan

As shown in Figure 1a, the same empathy score (a score of 1.0 here means the lowest empathy) can be self-reported for essays with clearly different empathic content, resulting in incorrect or misleading labels. Such noise may confuse regression models, as they are penalised for producing predictions aligned with textual meaning when the corresponding label is unreliable. Without explicitly accounting for this discrepancy, models risk overfitting to noisy supervision and producing incorrect empathy scores. To address this, our method estimates uncertainty to capture samples with label noise (Figure 1b).

Robust learning with noisy labels has been extensively studied in the literature for classification problems, but it is rarely explored for regression problems [7]. Most robust learning with noisy label algorithms, e.g., Zheng et al. [8] and Ye et al. [9], leverage class probabilities. Regression and classification modelling are fundamentally different due to the type of prediction (class probability versus real numbers) and subsequently the type of cost function (cross-entropy versus error-based loss formulation). Due to the above conflicts in modelling, these robust learning algorithms designed for classification do not apply to regression problems [7, 10]. Our proposed method, UPLME, is specifically designed for textual regression problems and has been demonstrated on the emerging topic of empathy regression.

We model label noise as heteroscedastic Uncertainty Quantification (UQ). UQ, an established concept, is broadly categorised into epistemic and aleatoric uncertainty [11, 12]. Epistemic uncertainty reflects uncertainty in the model parameters due to limited data, while aleatoric uncertainty captures the inherent noise in the data itself [11]. Aleatoric uncertainty is further divided into homoscedastic (constant across inputs) and heteroscedastic (input-dependent) uncertainty. Since label noise, by definition, varies across input samples, we capture label noise as heteroscedastic UQ. Following recent evidence in the UQ literature on dissolving the dichotomy of epistemic and aleatoric uncertainty, we model the *source* of uncertainty, more specifically, annotation-related uncertainty [13].

To ensure a higher uncertainty score on noisy labels and a lower uncertainty score on clean labels (Figure 1), we propose a novel penalty loss that penalises degenerate uncertainty scores (i.e., low uncertainty on noisy samples and high uncertainty on clean samples). Theoretically, empathy is dependent on how similar the pair of input texts for which empathy is being predicted, so we further propose a novel alignment loss to enforce the similarity on the representation space of the input pairs.

UQ itself is a challenging task due to the lack of any ground truth uncertainty score for supervision [14]. Like robust learning with noisy labels, UQ has also been less explored in regression domains due to the challenges of the output label space being continuous rather than discrete class probabilities [7]. We propose a probabilistic fine-tuning of Pre-trained Language Models (PLMs) to estimate uncertainty using the concept of Bayesian Neural Networks (BNNs). Our key contributions include:

- 1) A novel penalty loss component to penalise degenerate uncertainty quantification related to label noise.

- 2) A novel alignment loss component to enforce a theoretical concept of empathy.
- 3) Uncertainty-aware probabilistic fine-tuning of PLM to capture label noise.
- 4) State-of-the-art empathy detection performance on two public benchmarks.

II. RELATED WORK & OUR NOVELTY

A. Empathy Detection

There has been a growing number of empathy detection studies over the past decade, with a significantly higher number of works on Natural Language Processing (NLP) types [2]. We refer interested readers to a recent systematic literature review on empathy detection [2].

Task formulations for empathy detection in NLP include detecting empathy in individuals through their written essays [15, 16], as well as empathic similarity between two narrative stories [17], among other scenarios. From a methodological point of view, almost all of these studies fine-tune Pre-trained Language Models (PLM), especially RoBERTa, as the backbone model [2]. Hasan et al. [6] addressed label noise in empathy detection as a *data-centric* approach. They re-annotated crowdsourced, noisy labels using a Large Language Model (LLM) and then selected an appropriate label between the two for each sample. Their empathy prediction model remained the same as those commonly found in the literature – fine-tuning the RoBERTa PLM. In contrast to existing empathy detection methods in the literature, our proposed model is probabilistic, predicting both empathy and uncertainty on each prediction with two novel loss components. Our method does not require the use of any LLM and still outperforms the prior work [6] on addressing label noise in empathy detection.

B. Robust Regression with Noisy Label

Among the few works on robust regression with noisy labels, Ouyang et al. [18] proposed an ensemble-based outlier detection method by combining seven established algorithms. Since different detectors can capture different aspects of noisy data, such aggregation was able to identify the most obvious outliers agreed by the ensemble. However, combining seven algorithms increases computational complexity, and importantly, their method is demonstrated only on numerical regression tasks. In the textual regression domain, Wang et al. [7] evaluated several external regularisation strategies (e.g., dropout and weight decay) and proposed an iterative algorithm to identify noisy labels. Unlike this line of work, we propose probabilistic fine-tuning of pre-trained language models that automatically downweights the contribution of noisy labels during training.

C. Uncertainty Quantification

Shukla et al. [14] introduced a self-supervised approach to estimate label uncertainty using covariance, where pseudo-labels are derived through a neighbourhood-based heuristic. To compare multivariate normal distributions, they reported that the 2-Wasserstein distance performs better than the KL

divergence. Recently, Zulfiqar et al. [12] uses Monte Carlo dropout (MCD) to capture uncertainty and a Gaussian mixture model to cluster high-confidence pseudo labels for positive and unlabelled learning problems. Liang et al. [19] proposed R-Drop, a regularisation method that minimises the divergence between outputs from two forward passes with dropout. This technique treats the two passes as stochastic sub-models within a single network. Their final loss combines negative log-likelihood with bidirectional KL divergence between the output distributions. However, R-Drop uses softmax probabilities, which makes it unsuitable for regression tasks.

Wang et al. [20] explored UQ in textual regression settings, where they used a threshold, as a hyperparameter, on the predicted uncertainty to filter less confident samples. Contrary to their work, our method does not require any such thresholds. Finally, Dai et al. [10] proposed UCVME, a semi-supervised learning framework based on variational model ensembling and MCD as a BNN approach. UC in the acronym UCVME stands for uncertainty consistency, as their framework imposes consistency on the estimated uncertainties of two separate models. Their UQ framework is designed and demonstrated for computer vision problems with ResNet as the backbone model. In contrast, our method is designed and demonstrated for textual regression tasks with PLM as the backbone. It does not require two separate models or consistency between models, yet it outperforms their approach.

III. METHOD

A. Problem Formulation

Let x_1 and x_2 be a pair of input texts and $y \in \mathbb{R}$ the corresponding empathy score. In the EmpStories dataset [17], y measures empathic similarity between two story texts. In the NewsEmp dataset [21], it refers to the degree of empathy expressed in essays written by study participants after reading newspaper articles.

We focus on *state* empathy in written narratives, which depends on both source and target contexts [22]. Accordingly, we adopt a *cross-encoder* formulation: we form the input sequence

$$z = [\text{CLS}, x_1, \text{SEP}, \text{SEP}, x_2, \text{SEP}], \quad (1)$$

tokenise it jointly, and feed it into a shared encoder $\mathcal{F}_\theta$ ¹. While a bi-encoder (two separate encoders) is possible, the cross-encoder can directly capture interactions between x_1 and x_2 .

B. Uncertainty-Aware Probabilistic Modelling

To capture both the predicted empathy score and the input-dependent uncertainty, we extend a PLM with parallel regression heads. This design enables the model to jointly learn the target value and its corresponding variance, which is the foundation of our probabilistic framework. Mathematically, we

augment $\mathcal{F}_\theta$ (e.g., RoBERTa, DeBERTa or ModernBERT) with two parallel regression heads:

$$\begin{aligned} h &= \mathcal{F}_\theta(z), \quad \hat{y} = w_m^\top h + b_m, \\ v &= w_v^\top h + b_v, \quad \sigma^2 = \text{softplus}(v) + \varepsilon, \end{aligned} \quad (2)$$

where $\hat{y}$ is the predicted mean, σ^2 is the aleatoric variance (clamped by $\varepsilon = 10^{-8}$ for stability), and h denotes the pooled sequence representation produced by the encoder, obtained by using the standard RoBERTa pooler (a linear layer followed by Tanh activation function) for RoBERTa and standard CLS pooling for other PLMs.

C. Variational Model Ensembling

We capture epistemic uncertainty using Monte Carlo dropout (MCD), which provides a variational Bayesian approximation without requiring a fully Bayesian neural network. In MCD, dropout is activated at inference time to simulate sampling from a posterior over models [10]. In particular, UPLME involves T stochastic forward passes through the probabilistic language model $\mathcal{F}_\theta$ during training, validation and test phases. We collect the set $\{\hat{y}_i^{(t)}, \sigma_i^{2,(t)}, h_i^{(t)}\}_{t=1}^T$, representing the predicted empathy scores, variances and intermediate representations from each stochastic pass. We then compute the ensemble means:

$$\begin{aligned} \bar{y}_i &= \frac{1}{T} \sum_{t=1}^T \hat{y}_i^{(t)}, \quad \bar{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^T \sigma_i^{2,(t)}, \\ \bar{h}_i &= \frac{1}{T} \sum_{t=1}^T h_i^{(t)}. \end{aligned} \quad (3)$$

We use a single probabilistic PLM, which has three loss components, as described in the following subsections.

1) *Negative Log-Likelihood*: A Negative Log-Likelihood (NLL) objective can adjust the influence of samples according to the variance parameters, which is equivalent to the estimated noise. Following Seitzer et al. [11], we use the improved version of it, a β -NLL loss, which down-weights high-variance points as:

$$\mathcal{L}_{\beta\text{-nll}} = \frac{1}{2} \sum_{i=1}^N \bar{\sigma}_i^{2\beta} \left(\log \bar{\sigma}_i^2 + \frac{(y_i - \bar{y}_i)^2}{\bar{\sigma}_i^2} \right), \quad (4)$$

where $\beta > 0$ controls the degree of variance rescaling, which is set to 0.5, following Seitzer et al. [11].

2) *Variance Penalty*: To discourage degenerate variance predictions (i.e., high uncertainty at low error and low uncertainty at high error), we apply an error-weighted penalty to the predicted variance. Specifically, we scale each variance by an exponentially decaying function of its squared error, and take the L2 norm across the batch:

$$\mathcal{L}_{\text{pen}} = \frac{1}{N} \left\| e^{-\alpha(y - \bar{y})^2} \times \bar{\sigma}^2 \right\|_2, \quad (5)$$

where $\alpha > 0$ modulates how sharply low-error points are penalised for high uncertainty. This formulation encourages the model to reserve high variance for genuinely uncertain samples, and therefore improves aleatoric UQ.

¹Note that while RoBERTa tokeniser has this formulation of two SEP tokens in separating x_1 and x_2 , DeBERTa and ModernBERT tokenisers assign only one SEP token in separating the inputs.

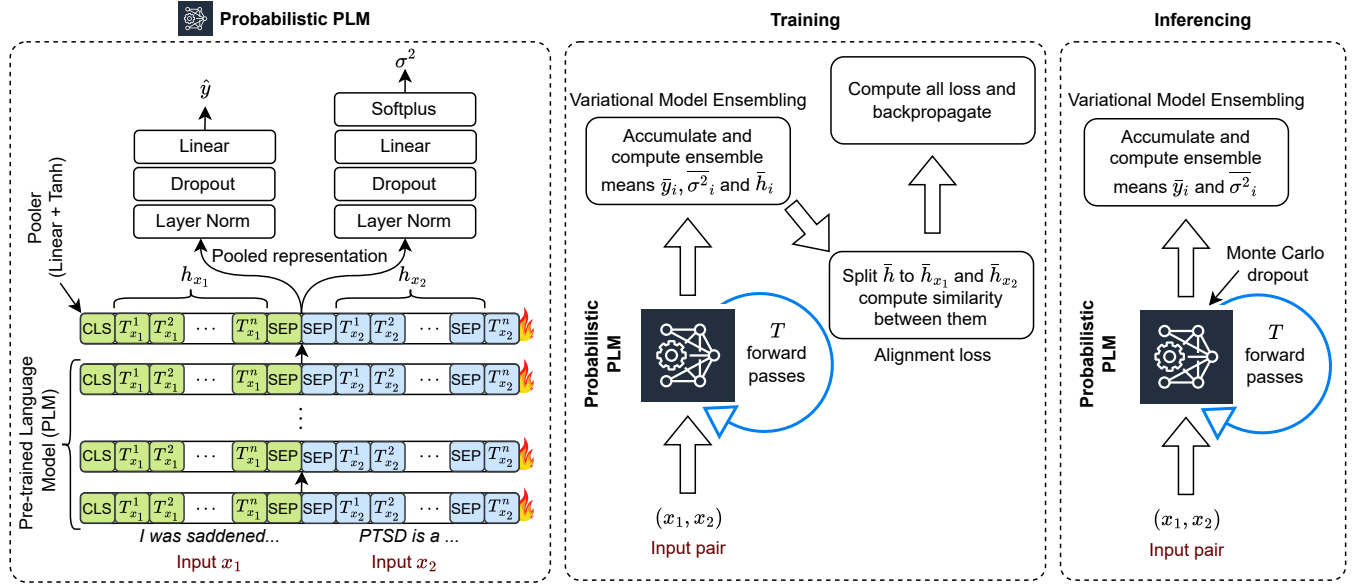


Fig. 2. Overview of our proposed UPLME framework. Input pairs are first concatenated and fed to a probabilistic Pre-trained Language Model (PLM) that predicts both empathy scores and heteroscedastic uncertainty. Uncertainty quantification is stabilised through variational model ensembling that includes multiple forward passes through the same model.

Algorithm 1 UPLME Training and Evaluation

Require: Batch $\mathcal{B} = \{(x_1^{(i)}, x_2^{(i)}, y^{(i)})\}_{i=1}^N$, encoder $\mathcal{F}_\theta$, heads $(w_m, b_m), (w_v, b_v)$, hyperparameters $\alpha, \lambda_1, \lambda_2$, passes T

Ensure: Total loss $\mathcal{L}_{\text{total}}$

- 1: $\triangleright$ *Training using Backpropagation* $\triangleleft$
- 2: **for** each $(x_1, x_2, y) \in \mathcal{B}$ **do**
- 3: Initialise $\sum \hat{y} \leftarrow 0, \sum \sigma^2 \leftarrow 0, \sum h_z \leftarrow 0$
- 4: **for** $t \leftarrow 1$ **to** T **do**
- 5: Concatenate x_1 and x_2 using Equation (1)
- 6: Forward pass through the probabilistic PLM using Equation (2)
- 7: Accumulate: $\sum \hat{y} += \hat{y}^{(t)}, \sum \sigma^2 += \sigma^{2,(t)}, \sum h_z += h_z^{(t)}$
- 8: Compute ensemble means using Equation (3)
- 9: Compute $\mathcal{L}_{\beta\text{-nll}}$ using Equation (4)
- 10: Compute $\mathcal{L}_{\text{pen}}$ using Equation (5)
- 11: Split h_z into h_{x_1}, h_{x_2} , and calculate similarity using Equation (6)
- 12: Compute $\mathcal{L}_{\text{align}}$ using Equation (7)
- 13: Calculate $\mathcal{L}_{\text{total}}$ using Equation (8) and accumulate for backpropagation.
- 14: $\triangleright$ *Evaluation* $\triangleleft$
- 15: **for** each (x_1, x_2) in validation/test set **do**
- 16: Activate dropout in $\mathcal{F}_\theta$ $\triangleright$ *Monte Carlo Dropout (MCD)*
- 17: Repeat Line 3 to Line 8 to compute $\bar{y}, \bar{\sigma}^2$
- 18: Use $\bar{y}, \bar{\sigma}^2$ for downstream metrics

3) *Alignment Loss*: The degree of alignment between two texts (e.g., a news article and its corresponding essay) should reflect the degree of empathy. To capture this, we compute the *cosine similarity* between their token representations.

the token embeddings for each segment:

$$\bar{h}_{x_1} = \frac{1}{|T_{x_1}|} \sum_{t \in T_{x_1}} \bar{h}_t, \quad \bar{h}_{x_2} = \frac{1}{|T_{x_2}|} \sum_{t \in T_{x_2}} \bar{h}_t, \quad (6)$$

$$s = \cos(\bar{h}_{x_1}, \bar{h}_{x_2}),$$

Given the input sequence defined in Equation (1), the encoder produces a contextualised representation $\bar{h}_z \in \mathbb{R}^{L \times d}$ for the full token sequence z . We extract individual representations $\bar{h}_{x_1}$ and $\bar{h}_{x_2}$ by splitting $\bar{h}_z$ at the SEP tokens and averaging

where T_{x_1} and T_{x_2} index the token positions of x_1 and x_2 within z . The similarity score $s \in [-1, 1]$ captures the degree of alignment between the two texts. To use this as a learning signal, we first rescale the original empathy labels (e.g., $[1, 7]$) to $[-1, 1]$ to match the scale of s . Let y' denote

the transformed label. We then define the alignment loss as

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^N (s_i - y'_i)^2. \quad (7)$$

This objective encourages the model to align the similarity between two texts with their associated empathy scores.

The overall loss for a *single probabilistic model* is, therefore:

$$\mathcal{L} = \mathcal{L}_{\beta\text{-nll}} + \lambda_1 \mathcal{L}_{\text{pen}} + \lambda_2 \mathcal{L}_{\text{align}}, \quad (8)$$

with weights $\lambda_{1,2}$ to be tuned as hyperparameters.

Figure 2 provides an overview of the complete UPLME framework. On the left, the probabilistic PLM is shown: input pairs (x_1, x_2) are concatenated and jointly encoded, with a pooled representation passed through two regression heads to predict the empathy score $\hat{y}$ and heteroscedastic variance σ^2 , respectively. During the training phase (middle), UPLME performs T stochastic forward passes and accumulates the predicted means, the variances, and the hidden representations for loss computations. The loss function consists of three components: the likelihood-based objective, the variance penalty, and the alignment loss between the representations of x_1 and x_2 . During inference (right), the same procedure is repeated with MCD to approximate posterior sampling, and ensemble averages of the predictions are used as final outputs. Overall training and evaluation steps of UPLME are shown in Algorithm 1.

D. UPLME vs Others

UPLME differs substantially from prior uncertainty-aware regression approaches in both design and training. Unlike UCVME [10], which co-trains two ResNet-based models with an uncertainty consistency loss in a semi-supervised setting for computer vision tasks, UPLME utilises a single PLM backbone for textual regression and achieves heteroscedastic uncertainty estimation without requiring dual models or explicit consistency constraints. Moreover, previous methods for robust regression under label noise often rely on external heuristics to filter or correct noisy data – for example, combining multiple outlier detectors into an ensemble [18] or iteratively removing suspected noisy labels in textual datasets [7] – and some require manual uncertainty thresholds to discard low-confidence predictions [20]. In contrast, UPLME directly learns a probabilistic output distribution and treats label noise as input-dependent (heteroscedastic) uncertainty, automatically down-weighting the influence of noisy examples via a β -NLL loss term instead of ad-hoc filtering. In addition to excelling in the UQ task, UPLME introduces two novel loss components: a variance penalty that discourages miscalibrated uncertainty estimates, and an alignment loss that reflects the semantic similarity of input pairs.

In terms of empathy detection workflows, prior works (e.g., [6, 16, 23] (Table II, to be presented later) primarily fine-tuned PLMs with their default mean squared error objective function. In contrast, UPLME comprises penalty loss, alignment loss, and variational model ensembling within a UQ framework.

TABLE I
STATISTICS OF THE DATASETS USED TO VALIDATE UPLME.

Name	# Train	# Validation	# Test	# Total
NewsEmp21 [24]	1,860	270	525	2,655
NewsEmp24 [16]	1,000	63	83	1,146
EmpStories [17]	1,500	100	400	4,000

IV. EXPERIMENTS AND RESULTS

A. Dataset & Preprocessing

We validate UPLME across three datasets: two different datasets from the NewsEmp series [16, 24] and the EmpStories dataset [17]. Table I presents statistics of the three datasets.

1) *NewsEmp Datasets*: Buechel et al. [21] pioneered the study of how people empathise with stories of suffering related to different entities (e.g., humans, animals and environment) portrayed in newspaper articles. After reading the articles, crowdsourced participants self-assessed their *state* empathy through Batson’s empathy scale [4] and subsequently wrote 300–800 character essays about their thoughts and feelings towards the news. The empathy label in this dataset is a continuous score ranging from 1.0 to 7.0, which is derived from the essay writers’ self-assessments. Multiple versions of this original dataset have been utilised in empathy detection challenges under the “Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)” [16]. Among these, the 2021 and 2024 releases, hereinafter referred to as the NewsEmp21 and NewsEmp24 datasets, are distinct datasets with no overlapping samples and are therefore used in this work. However, some samples from the first 2018 release [21] and some samples from the 2023 release [25] overlap with the NewsEmp21 and NewsEmp24 datasets, respectively, and are therefore not used in this work.

Our motivation behind choosing the NewsEmp datasets is two-fold. First, these datasets are the most widely used public benchmarks across all empathy benchmarks in the literature [2], allowing us to compare our method’s performance across a broader range of works. Second, prior work [6] demonstrated label noise in these datasets, supported by evidence of noise and prior literature suggesting that crowdsourcing often results in incorrect data [5, 26, 27, 28]. Having such label noise, the NewsEmp datasets fit well with the goal of UPLME – robust empathy regression under the label noise problem.

The news articles in the NewsEmp datasets are long sequences, with a maximum length of 20,047 characters. To accommodate these articles in the typical PLMs’ (e.g., RoBERTa) context length, we use the summarised articles reported in [23], which makes the maximum length only 987 characters.

Following Giorgi et al. [16]’s pioneering work on the NewsEmp24 dataset, we train models using the combined training sets of multiple releases. Specifically, we combine the training set of the latest 2024 release with the training and development sets of its 2021 release. We further augment the training data using WordNet-based synonym substitution, replacing 10% of the words in each training sample. This augmentation doubles the size of the training set and intro-

duces lexical diversity while preserving original semantics. This resulted in a total of 6260 training samples, while the validation set (63 samples) remains the same as the original 2024 release [16]. We evaluate the same trained model on both the NewsEmp21 and NewsEmp24 test sets, which contain 525 and 83 samples, respectively.

2) *EmpStories Dataset*: Shen et al. [17] recently released pairs of narrative stories and corresponding empathic similarity. Two human annotators rate empathic similarity on a 4-point Likert scale, which is later averaged, resulting in the final empathy labels as $\{1.0, 1.5, \dots, 3.5, 4.0\}$.

Although we do not suggest any presence of label noise in the EmpStories dataset, we use this dataset to demonstrate the effectiveness of UPLME on a different type of dataset apart from NewsEmp. Key difference between the NewsEmp and the EmpStories includes: (1) the former (NewsEmp) includes newspaper articles and written essays, while the latter (EmpStories) includes a pair of similar personal narrative stories, and (2) the former is annotated through *self-assessment* by the essay writers themselves, while the latter is annotated by two third-party annotators.

B. Evaluation Protocol

Pearson Correlation Coefficient (PCC) is the metric adopted by the WASSA empathy detection challenge for the NewsEmp datasets. While PCC capture the linear relationship between the prediction and the ground truth, it fails to capture the magnitude of prediction error. Therefore, in addition to PCC, we report Concordance Correlation Coefficient (CCC) and Root Mean Square Error (RMSE). CCC takes into account both the linear relationship and the magnitude of the error, whereas RMSE solely captures the magnitude of the error.

To evaluate the quality UQ, we adopt all three UQ metrics proposed by Wang et al. [20]: Calibration error (CAL), Sharpness (SHP) and Negative Log-Probability Density (NLPD).

The training, validation and test splits are already predefined in all three datasets we used. Note that the ground-truth labels for the NewsEmp21 test set are not publicly released; evaluation is only available through the CodaLab challenge platform². We run each model using three different initialisations, set by random seeds of 0, 42, and 100, and compare based on median and peak (i.e., the best) scores of the models across these three seeds.

C. Implementation Details

UPLME was primarily implemented using PyTorch 2.2.0 with ROCm 5.7.3 in Python 3.10 and also tested on PyTorch 2.7.1 with ROCm 6.3.3 in Python 3.12.3. Deterministic training and evaluation are ensured using the PyTorch Lightning package. All experiments were conducted on a single AMD Instinct™ MI250X GPU (64 GB) running on SUSE Linux Enterprise Server 15 SP5.

Pre-trained checkpoints of RoBERTa (version: roberta-base) [44], DeBERTa (version:

²<https://codalab.lisn.upsaclay.fr/competitions/834>, noting that the same dataset has been used both in the WASSA 2021 [24] and the 2022 [15] challenges.

TABLE II
COMPARISON OF OUR PROPOSED MODEL’S PERFORMANCE WITH THE PERFORMANCE *reported in the literature*. A SINGLE ROBERTA MODEL UNDER THE UPLME FRAMEWORK, OPTIMISED FOR THE NEWSEMP24 VALIDATION SET, IS TESTED ON BOTH THE NEWSEMP21 AND NEWSEMP24 TEST SETS.

Reference	Venue & Year	Model	PCC ↑
<i>NewsEmp21</i>			
Butala et al. [29]	ACL WASSA 2021	BERT-MLP	0.358
Vasava et al. [30]	ACL WASSA 2022	RoBERTa-MLP	0.470
Ghosh et al. [31]	ACL WASSA 2022	BERT-MLP	0.479
Qian et al. [32]	ACL WASSA 2022	RoBERTa	0.504
Hasan et al. [6]	EACL 2024	RoBERTa-MLP	0.505
Vettigli and Sorgente [33]	EACL WASSA 2021	LR	0.516
Kulkarni et al. [34]	EACL WASSA 2021	RoBERTa-MLP	0.517
Lahnala et al. [35]	ACL WASSA 2022	RoBERTa	0.524
Chen et al. [36]	ACL WASSA 2022	RoBERTa	0.537
Plaza-del-Arco et al. [37]	ACL WASSA 2022	RoBERTa	0.541
Mundra et al. [38]	EACL WASSA 2021	ELECTRA + RoBERTa	0.558
Ours	–	UPLME (RoBERTa)	0.580
<i>NewsEmp24</i>			
Numanoglu et al. [39]	ACL WASSA 2024	BERT	0.290
Chevi and Aji [40]	ACL WASSA 2024	MLP	0.345
Frick and Steinebach [41]	ACL WASSA 2024	RoBERTa	0.375
Pereira et al. [42]	ACL WASSA 2024	<i>Not reported</i>	0.390
Li et al. [43]	ACL WASSA 2024	Llama 3 8B	0.474
Giorgi et al. [16]	ACL WASSA 2024	RoBERTa	0.629
Ours	–	UPLME (RoBERTa)	0.634

microsoft/deberta-v3-base) [45], and ModernBERT (version: answerdotai/ModernBERT-base) [46] are accessed from HuggingFace. We use the TextAttack package [47] for training data augmentation.

Automated tuning of loss weights was conducted using the Optuna hyperparameter tuning framework [48], with the median pruner (warmup steps = 5) and the default “tree-structured Parzen estimator” as the sampler. Tuning was conducted between $\{1.0, 1.5\}$ for α and $[0, 50]$ with a step of 0.1 for both λ_1 and λ_2 , with the objective of selecting the final set of parameters that yields the best validation CCC score. We tuned α , λ_1 and λ_2 of UPLME and w (factor weighing the contribution of two models’ output consistency) of UCVME [10].

Other hyperparameters were kept fixed throughout the experiment, with most of them being selected from the literature. For example, following the pioneering work [44] releasing RoBERTa, we set $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e-6$ and weight decay = 0.1 of the AdamW optimiser and train (fine-tune) with a learning rate of $3e-5$. DeBERTa was fine-tuned with a learning rate of $3e-5$ and weight decay of 0.01, as reported in [45]. ModernBERT was fine-tuned with the same learning rate and weight decay as RoBERTa. Training was conducted for a maximum of 6,000 steps with a batch size of 16 for both training and evaluation. We used a linear learning rate scheduler with a warmup period of the first 3% of the training steps. We employ a delayed-start early-stopping strategy, which ensures a minimum of 5 epochs of training, followed by early stopping based on the validation CCC score.

D. Main Results

We benchmark the performance of UPLME against the State-Of-The-Art (SOTA) performance *reported in the literature* (Table II) and also against recent baseline methods implemented by us (Table III). Since RoBERTa is the most

TABLE III

UPLME VERSUS OUR IMPLEMENTATION OF BASELINE METHODS, TRAINED AND EVALUATED ON THE NEWSEMP24 AND THE EMPSTORIES DATASETS (THESE METRICS ON THE NEWSEMP21 DATASET WERE NOT CALCULATED DUE TO THE UNAVAILABILITY OF GROUND TRUTH TEST LABELS). UCVME IS A RECENT UNCERTAINTY-AWARE VARIATIONAL ENSEMBLING-BASED APPROACH.

Framework	PCC $\uparrow$	CCC $\uparrow$	SCC $\uparrow$	RMSE $\downarrow$	CAL $\downarrow$	SHP $\downarrow$	NLPD $\downarrow$
<i>NewsEmp24</i>							
PLM-MLP	0.554(0.569)	0.476(0.499)	0.453(0.488)	1.384(1.368)			
UCVME [10]	0.587(0.589)	0.510(0.530)	0.493(0.496)	1.337(1.317)	0.571(0.582)	0.001(0.001)***	2085042.875(2449522.5)
UPLME (Ours)	0.623(0.634)	0.523(0.534)	0.503(0.53)	1.322(1.313)	0.376(0.402)**	0.268(0.291)	10.775(14.718)***
<i>EmpStories</i>							
PLM-MLP	0.180(0.215)	0.113(0.150)	0.168(0.212)	0.755(0.745)			
UCVME [10]	0.239(0.250)	0.180(0.189)	0.244(0.244)	0.748(0.741)	0.534(0.566)	0.002(0.003)***	94886.258(142711.703)
UPLME (Ours)	0.256(0.301)	0.217(0.237)	0.250(0.292)	0.747(0.722)	0.262(0.289)***	0.186(0.232)	6.759(7.477)*

Reported metrics are in *median(peak)* format, calculated over three random initialisations.

Boldface scores indicate the best scores.

Asterisk (*) refer to the statistical significance level from an independent t-test between UCVME and UPLME (* means $0.01 < p\text{-value} \leq 0.05$; ** means $0.001 < p\text{-value} \leq 0.01$; *** means $0.0001 < p\text{-value} \leq 0.001$)

commonly used backbone for empathy regression across the datasets we experimented [2], we primarily report results using RoBERTa unless specified otherwise.

1) *Benchmarking Results:* Since the literature on the NewsEmp datasets primarily uses the PCC evaluation metric, Table II compares our method’s performance with the performance reported in the literature in terms of the PCC score. On both datasets, UPLME outperforms all prior results.

The prior SOTA work [38] on the NewsEmp21 dataset employs an ensemble of ELECTRA and RoBERTa, reporting a PCC of 0.558, which is now surpassed by our proposed UPLME framework using a single RoBERTa PLM. On the NewsEmp24 dataset, our method also provides the best PCC score, even outperforming Llama 3 LLM used in [43].

For the NewsEmp21 dataset, we do not retrain the model; instead, we use the same model that was optimised for the NewsEmp24 dataset. Even so, UPLME outperforms other studies on the NewsEmp21 dataset, which suggests that UPLME generalises well.

2) *Benchmarking Methods:* In addition to comparing our results with those reported in the literature, we also compare UPLME with two baseline methods. PLM-MLP is UPLME without any UQ and so uses a mean squared error loss function, similar to many methods (e.g., [6, 16, 23]) presented on Table II. UCVME is a UQ method that utilises variational model ensembling, but it requires two models to enforce consistency in the uncertainty prediction. Although originally proposed for computer vision tasks using ResNet, we apply the same modelling technique (two models and their loss functions) to two RoBERTa PLMs. As presented in Table III, our method outperforms both PLM-MLP and UCVME in terms of regression metrics (PCC, CCC, SCC and RMSE).

In terms of UQ metrics, our method outperforms UCVME in CAL and NLPD metrics. The work [20] proposing these UQ metrics noted that when a model achieves the best SHP but has a very large NLPD (as is the case here for UCVME), we should compare models based on PCC, followed by CAL and NLPD. We can, therefore, conclude that our proposed method, UPLME, outperforms UCVME in the UQ task as well.

E. Ablation Study

To better understand the contribution of each component and hyperparameters of UPLME, we analyse model performance on both NewsEmp24 and EmpStories test sets under different configurations as reported in Table IV. As demonstrated in *Theme A*, when both penalty and alignment losses are removed ($\lambda_{1,2} = 0$), performance drops substantially across most metrics in both datasets. Adding only the penalty loss ($\lambda_1 = 1, \lambda_2 = 0$) yields the largest improvement. The alignment loss alone ($\lambda_1 = 0, \lambda_2 = 1$) improves calibration (lower median CAL) but contributes less to empathy prediction performance in the NewsEmp dataset. A plausible explanation is that, while alignment between articles and essays is theoretically grounded, each article is paired with multiple essays that often correspond to different empathy scores, which weakens the consistency of this assumption.

Theme B shows that, overall, variational model ensembling with $T = 4$ demonstrates better performance than a single forward pass with no ensembling ($T = 1$) and ensembling with $T = \{2, 3, 5\}$. In *Theme C*, data augmentation shows performance improvements over no augmentation in both datasets.

While we primarily demonstrated the effectiveness of UPLME using the RoBERTa PLM, Table IV *Theme D* shows that UPLME is independent of the specific PLM choice. RoBERTa is replaced by ModernBERT or DeBERTa, which work without any modification to the UPLME framework, suggesting that UPLME can be readily adapted to various transformer-based encoders. RoBERTa yields the highest scores, which supports the general consensus that RoBERTa has become the most popular PLM in the empathy computing literature [2].

F. Qualitative Analysis

Figure 3 evaluates whether the uncertainty scores estimated by UPLME behave as intended under controlled noise injection. For this demonstration, 30% of the training samples’ labels are shifted by 3.0 (e.g., $1.0 \rightarrow 4.0$) to create a noisy subset. First, the uncertainty distribution significantly differs between clean and noisy samples: noisy samples receive higher variance estimates, while clean samples concentrate around

TABLE IV
TEST SET PERFORMANCE OF UPLME WITH DIFFERENT MODEL CONFIGURATIONS ON THE NEWSEMP24 AND EMPSTORIES DATASETS.

Theme [†]	Configuration	PCC $\uparrow$	CCC $\uparrow$	SCC $\uparrow$	RMSE $\downarrow$	CAL $\downarrow$	SHP $\downarrow$	NLPD $\downarrow$
A _{NewsEmp}	$\lambda_{1,2} = 0$	0.580(0.597)	0.514(0.516)	0.483(0.503)	1.356(1.318)	0.362(0.393)	0.270(0.281)	11.555(14.63)
	$\lambda_1 = 1, \lambda_2 = 0$	0.586(0.610)	0.512(0.529)	0.513(0.526)	1.365(1.340)	0.318(0.323)	0.310(0.318)	8.668(11.514)
	$\lambda_1 = 0, \lambda_2 = 1$	0.552(0.563)	0.489(0.517)	0.466(0.468)	1.386(1.351)	0.292(0.339)	0.279(0.317)	10.630(11.966)
	$\lambda_{1,2} = 1$	0.551(0.579)	0.486(0.499)	0.461(0.487)	1.404(1.39)	0.369(0.388)	0.268(0.343)	12.498(16.858)
A _{EmpStories}	$\lambda_{1,2} = 0$	0.226(0.262)	0.185(0.188)	0.220(0.267)	0.755(0.740)	0.232(0.285)	0.269(0.326)	2.839(4.466)
	$\lambda_1 = 1, \lambda_2 = 0$	0.254(0.272)	0.203(0.212)	0.252(0.270)	0.747(0.737)	0.242(0.265)	0.230(0.298)	3.784(3.955)
	$\lambda_1 = 0, \lambda_2 = 1$	0.204(0.303)	0.158(0.249)	0.215(0.302)	0.784(0.782)	0.208(0.274)	0.311(0.353)	2.021(2.596)
	$\lambda_{1,2} = 1$	0.231(0.239)	0.197(0.197)	0.225(0.228)	0.759(0.741)	0.168(0.185)	0.32(0.333)	1.773(2.347)
B _{NewsEmp}	$T = 1$	0.568(0.593)	0.509(0.518)	0.474(0.494)	1.369(1.337)	0.359(0.416)	0.279(0.301)	11.948(15.999)
	$T = 2$	0.556(0.559)	0.493(0.503)	0.479(0.481)	1.376(1.364)	0.324(0.379)	0.257(0.280)	13.598(15.349)
	$T = 3$	0.572(0.584)	0.486(0.530)	0.482(0.485)	1.376(1.328)	0.399(0.403)	0.240(0.241)	15.289(17.327)
	$T = 4$	0.578(0.584)	0.524(0.533)	0.488(0.496)	1.329(1.314)	0.336(0.376)	0.260(0.293)	11.976(13.819)
	$T = 5$	0.578(0.579)	0.518(0.521)	0.466(0.467)	1.342(1.328)	0.356(0.396)	0.239(0.264)	13.984(16.252)
B _{EmpStories}	$T = 1$	0.283(0.283)	0.235(0.237)	0.288(0.291)	0.740(0.735)	0.243(0.268)	0.204(0.207)	5.088(5.259)
	$T = 2$	0.280(0.289)	0.232(0.247)	0.279(0.280)	0.743(0.735)	0.236(0.270)	0.222(0.231)	4.501(5.761)
	$T = 3$	0.251(0.311)	0.220(0.259)	0.240(0.302)	0.763(0.725)	0.203(0.289)	0.232(0.243)	4.162(7.574)
	$T = 4$	0.289(0.308)	0.245(0.258)	0.279(0.298)	0.759(0.722)	0.226(0.244)	0.229(0.232)	4.153(6.616)
	$T = 5$	0.263(0.269)	0.220(0.233)	0.248(0.284)	0.751(0.736)	0.264(0.274)	0.210(0.220)	4.930(7.666)
C _{NewsEmp}	Augmentation	0.578(0.584)	0.524(0.533)	0.488(0.496)	1.329(1.314)	0.336(0.376)	0.260(0.293)	11.976(13.819)
	No augmentation	0.436(0.508)	0.354(0.426)	0.369(0.413)	1.439(1.367)	0.244(0.290)	0.489(0.547)	3.191(5.977)
C _{EmpStories}	Augmentation	0.289(0.308)	0.245(0.258)	0.279(0.298)	0.759(0.722)	0.226(0.244)	0.229(0.232)	4.153(6.616)
	No augmentation	0.249(0.301)	0.210(0.229)	0.269(0.296)	0.752(0.735)	0.226(0.246)	0.235(0.281)	3.56(4.417)
D _{NewsEmp}	RoBERTa	0.623(0.634)	0.523(0.534)	0.503(0.530)	1.322(1.313)	0.376(0.402)	0.268(0.291)	10.775(14.718)
	ModernBERT	0.527(0.562)	0.467(0.502)	0.467(0.482)	1.446(1.367)	0.386(0.461)	0.257(0.267)	15.505(20.052)
	DeBERTa	0.568(0.595)	0.473(0.510)	0.464(0.517)	1.388(1.383)	0.414(0.470)	0.227(0.255)	17.978(21.78)
D _{EmpStories}	RoBERTa	0.289(0.308)	0.245(0.258)	0.279(0.298)	0.759(0.722)	0.226(0.244)	0.229(0.232)	4.153(6.616)
	ModernBERT	0.186(0.254)	0.157(0.159)	0.179(0.25)	0.784(0.720)	0.259(0.324)	0.203(0.318)	6.041(10.204)
	DeBERTa	0.240(0.309)	0.211(0.248)	0.228(0.286)	0.757(0.717)	0.252(0.258)	0.198(0.234)	5.783(5.853)

[†] A – different loss components; B – variational ensembling; C – training data augmentation; D – different PLM encoders.

Reported metrics are in *median(peak)* format, calculated over three random initialisations.

Boldface scores indicate the best scores in each theme.

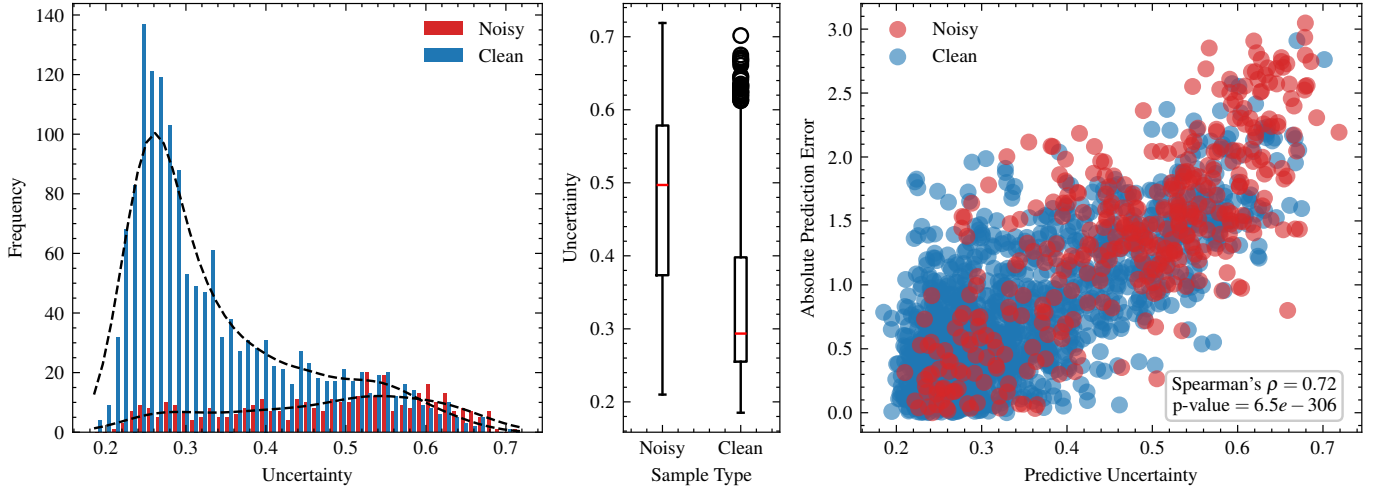


Fig. 3. Our proposed penalty loss component ensures predicted uncertainty captures label noise. **Left** and **Middle**: comparison of predictive uncertainty between noisy and clean samples, showing that UPLME estimates higher uncertainty for noisy samples. **Right**: relationship between absolute prediction error and estimated uncertainty on both noisy and clean subsets, showing strong statistically-significant positive correlation (Spearman’s $\rho = 0.72$).

lower uncertainty regions. Second, the predictive uncertainty strongly correlates with absolute prediction error (Spearman’s $\rho = 0.72$, with a p-value of $6.5e-306$), which implies that the estimated uncertainty captures label noise with a very high statistical significance.

Figure 4 examines how UPLME organises input contexts in its latent representation space. Both t-SNE and UMAP reveal a clear, continuous gradient in the embedding geometry that aligns with the ground-truth empathy scores, suggesting that UPLME has learned representations that correspond to the

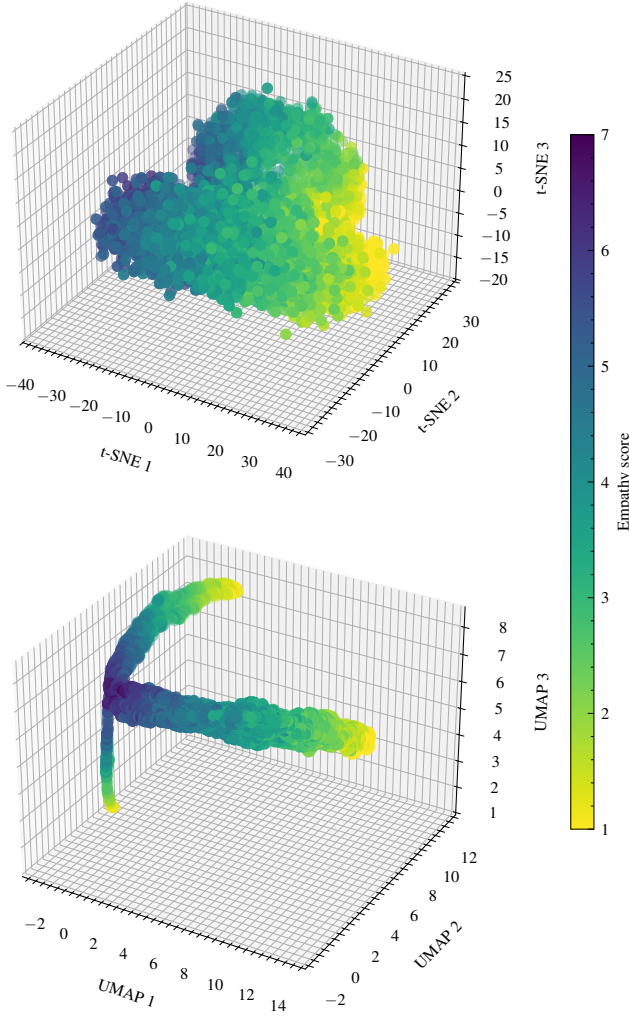


Fig. 4. 3D t-SNE (left) and UMAP (right) projections of UPLME’s learned representations on the NewsEmp training set. Each point corresponds to an input essay and is coloured by its ground-truth empathy score (1–7).

ordinal scale of empathy.

G. Limitations

While UPLME demonstrates robust performance, it has a few limitations. First, the variance penalty and alignment loss introduce additional hyperparameters whose optimal values may vary across datasets. Second, the use of variational ensembling increases computational cost relative to baselines with a single PLM, which may hinder scalability for larger model backbones or datasets. Nevertheless, the computational requirement is still smaller than that of LLMs or multiple ensembles of PLMs.

V. CONCLUSION

Noisy labels challenge empathy regression from self-reported annotation. This paper proposes UPLME, an uncertainty-aware probabilistic language modelling framework for empathy regression tasks from text sequences. UPLME estimates label noise-related uncertainty to down-weight the contribution from these noisy samples during training. The

proposed framework includes two novel loss components, one to stabilise UQ and another to enforce similarity on the input text pairs on which empathy is being detected. Experiments across three datasets show that UPLME is effective both in providing the SOTA empathy detection performance and in improving the UQ. In future, UPLME can be extended to semi-supervised learning frameworks, where the uncertainty estimation can improve pseudo-label quality.

ETHICS STATEMENT

This work uses publicly available empathy datasets collected through prior studies. All datasets were originally released with institutional ethics approval, and we followed their intended use for research purposes. Our method focuses on improving robustness to noisy self-reported labels and does not attempt to infer private or sensitive attributes such as demographic details. We acknowledge that automatic empathy detection may be misused, and emphasise that our framework is designed strictly for advancing methodological research in empathy and similar textual regression tasks.

ACKNOWLEDGMENTS

This work was supported by resources provided by the Pawsey Supercomputing Research Centre with funding from the Australian Government and the Government of Western Australia. We thank Dr Moloud Abdar, Senior Data Scientist at The University of Queensland, for sharing his opinion on uncertainty quantification during an early stage of this work.

REFERENCES

- [1] M. L. Hoffman, “Toward a theory of empathic arousal and development,” in *The Development of Affect*, M. Lewis and L. A. Rosenblum, Eds., Boston, MA: Springer US, 1978, pp. 227–256.
- [2] M. R. Hasan, M. Z. Hossain, S. Ghosh, A. Krishna, and T. Gedeon, “Empathy detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods,” *IEEE Transactions on Affective Computing*, pp. 1–19, 2025.
- [3] R. Banerjee and S. K. Pal, “Z*-numbers, data structures, and thinking in machine-mind architecture,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 5, pp. 686–695, 2020.
- [4] C. D. Batson, J. Fultz, and P. A. Schoenrade, “Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences,” *Journal of personality*, vol. 55, no. 1, pp. 19–39, 1987.
- [5] K. B. Sheehan, “Crowdsourcing research: Data collection with amazon’s mechanical turk,” *Communication Monographs*, vol. 85, no. 1, pp. 140–156, 2018.
- [6] M. R. Hasan, M. Z. Hossain, T. Gedeon, and S. Rahman, “LLM-GE: Large language model-guided prediction of people’s empathy levels towards newspaper article,” in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds., St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2215–2231.
- [7] Y. Wang, T. Baldwin, and K. Verspoor, “Noisy label regularisation for textual regression,” in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari et al., Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4228–4240.
- [8] S. Zheng, P. Wu, A. Goswami, M. Goswami, D. Metaxas, and C. Chen, “Error-bounded correction of noisy labels,” in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, Jul. 2020, pp. 11 447–11 457.
- [9] S. Ye, D. Chen, S. Han, and J. Liao, “Learning with noisy labels for robust point cloud segmentation,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6423–6432.

- [10] W. Dai, X. Li, and K.-T. Cheng, "Semi-supervised deep regression with uncertainty consistency and variational model ensembling via bayesian neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, Jun. 2023, pp. 7304–7313.
- [11] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," in *International Conference on Learning Representations (ICLR 2022)*, Apr. 2022.
- [12] M. I. Zulfikar, A. Khalid, L. Chen, and S. Dhelim, "Uncertainty-aware neighbor calibration for positive and unlabeled learning in large machine learning models," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–14, 2025.
- [13] M. Kirchhof, G. Kasneci, and E. Kasneci, "Reexamining the aleatoric and epistemic uncertainty dichotomy," in *The Fourth Blogpost Track at ICLR 2025*, 2025.
- [14] M. Shukla, A. Shameem, M. Salzmann, and A. Alahi, "Towards self-supervised covariance estimation in deep heteroscedastic regression," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
- [15] V. Barriere, S. Tafreshi, J. Sedoc, and S. Alqahtani, "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*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 214–227.
- [16] S. Giorgi, J. Sedoc, V. Barriere, and S. Tafreshi, "Findings of WASSA 2024 shared task on empathy and personality detection in interactions," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, J. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 369–379.
- [17] J. Shen, M. Sap, P. Colon-Hernandez, H. Park, and C. Breazeal, "Modeling empathic similarity in personal narratives," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6237–6252.
- [18] B. Ouyang, Y. Song, Y. Li, G. Sant, and M. Bauchy, "EBOD: An ensemble-based outlier detection algorithm for noisy datasets," *Knowledge-Based Systems*, vol. 231, p. 107400, 2021.
- [19] X. Liang et al., "R-Drop: Regularized dropout for neural networks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 10890–10905.
- [20] Y. Wang, D. Beck, T. Baldwin, and K. Verspoor, "Uncertainty estimation and reduction of pre-trained models for text regression," *Transactions of the Association for Computational Linguistics*, vol. 10, B. Roark and A. Nenkova, Eds., pp. 680–696, 2022.
- [21] S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc, "Modeling empathy and distress in reaction to news stories," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4758–4765.
- [22] M. R. Hasan, M. Z. Hossain, A. Krishna, S. Rahman, and T. Gedeon, "Are you really empathic? evidence from trait, state and speaker-perceived empathy, and physiological signals," *arXiv preprint arXiv:2509.16923*, 2025.
- [23] M. R. Hasan, M. Z. Hossain, T. Gedeon, S. Soon, and S. Rahman, "Curtin OCAI at WASSA 2023 empathy, emotion and personality shared task: Demographic-aware prediction using multiple transformers," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 536–541.
- [24] S. Tafreshi, O. De Clercq, V. Barriere, S. Buechel, J. Sedoc, and A. Balahur, "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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 92–104.
- [25] D. Omitaomu et al., "Empathic conversations: A multi-level dataset of contextualized conversations," *arXiv preprint arXiv:2205.12698*, 2022.
- [26] R. Jia, Z. R. Steelman, and B. H. Reich, "Using mechanical turk data in research: Risks, rewards, and recommendations," *Communications of the Association for Information Systems*, vol. 41, no. 1, p. 14, 2017.
- [27] J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon, "Detecting and deterring insufficient effort responding to surveys," *Journal of Business and Psychology*, vol. 27, pp. 99–114, 2012.
- [28] W. O'Brochta and S. Parikh, "Anomalous responses on amazon mechanical turk: An Indian perspective," *Research & Politics*, vol. 8, no. 2, 2021.
- [29] Y. Butala, K. Singh, A. Kumar, and S. Shrivastava, "Team Phoenix at WASSA 2021: Emotion analysis on news stories with pre-trained language models," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 274–280.
- [30] H. Vasava, P. Uikey, G. Wasnik, and R. Sharma, "Transformer-based architecture for empathy prediction and emotion classification," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 261–264.
- [31] S. Ghosh, D. Maurya, A. Ekbal, and P. Bhattacharyya, "Team IITP-AINLPM 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*, 2022, pp. 255–260.
- [32] S. Qian, C. Orašan, D. Kanojia, H. Saadany, and F. Do Carmo, "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*, 2022, pp. 271–275.
- [33] G. Vettigli and A. Sorgente, "EmpNa at WASSA 2021: A lightweight model for the prediction of empathy, distress and emotions from reactions to news stories," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 264–268.
- [34] A. Kulkarni, S. Somwase, S. Rajput, and M. Marathe, "PVG at WASSA 2021: A multi-input, multi-task, transformer-based architecture for empathy and distress prediction," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 105–111.
- [35] A. Lahkala, C. Welch, and L. Flek, "CAISA at WASSA 2022: Adapter-tuning for empathy prediction," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 280–285.
- [36] Y. Chen, Y. Ju, and S. Kübler, "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*, 2022.
- [37] F. M. Plaza-del-Arco, J. Collado-Montañez, L. A. Ureña, and M.-T. Martín-Valdivia, "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*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 239–244.
- [38] J. Mundra, R. Gupta, and S. Mukherjee, "WASSA@IITK at WASSA 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 112–116.
- [39] A. Numanoglu, S. Ateş, N. Cicekli, and D. Küçük, "Empathify at WASSA 2024 empathy and personality shared task: Contextualizing empathy with a BERT-based context-aware approach for empathy detection," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 393–398.
- [40] R. Chevi and A. Aji, "Daisy at WASSA 2024 empathy and personality shared task: A quick exploration on emotional pattern of empathy and distress," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 420–424.
- [41] R. Frick and M. Steinebach, "Fraunhofer SIT at WASSA 2024 empathy and personality shared task: Use of sentiment transformers and data augmentation with fuzzy labels to predict emotional reactions in

- conversations and essays,” in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 435–440.
- [42] P. Pereira, H. Moniz, and J. P. Carvalho, “ConText at WASSA 2024 empathy and personality shared task: History-dependent embedding utterance representations for empathy and emotion prediction in conversations,” in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 448–453.
- [43] T. Li, N. Rusnachenko, and H. Liang, “Chinchunmei at WASSA 2024 empathy and personality shared task: Boosting LLM’s prediction with role-play augmentation and contrastive reasoning calibration,” in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 385–392.
- [44] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [45] P. He, J. Gao, and W. Chen, “DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [46] B. Warner et al., *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*, 2024.
- [47] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, “TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., Association for Computational Linguistics, Oct. 2020, pp. 119–126.
- [48] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.