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

Raphael Antonius Frick and **Martin Steinebach**
Fraunhofer SIT | ATHENE Center
Rheinstraße 75, Darmstadt, Germany
{raphael.frick, martin.steinebach}@sit.fraunhofer.de

Abstract

Predicting emotions and emotional reactions during conversations and within texts poses challenges, even for advanced AI systems. The second iteration of the WASSA Empathy and Personality Shared Task focuses on creating innovative models that can anticipate emotional responses to news articles containing harmful content across four tasks. In this paper, we introduce our Fraunhofer SIT team's solutions for the three tasks: Task 1 (CONVD), Task 2 (CONVT), and Task 3 (EMP). It involves combining LLM-driven data augmentation with fuzzy labels and fine-tuning RoBERTa models pre-trained on sentiment classification tasks to solve the regression problems. In the competition, our solutions achieved 1st place in Track 1 (CONV-dialog), 8th in Track 2 (CONV-turn), and 3rd place in Track 3 (EMP).

1 Introduction

Consuming news articles and user-generated content online can evoke diverse emotions in individuals. Detecting empathic reactions to such content, often influenced by a reader's personality, remains a formidable challenge, even for advanced artificial intelligence (AI) systems.

The second iteration of the *Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (Giorgi et al., 2024) shared task focuses on creating AI models capable of predicting empathy, emotion, and personality. For this, akin to the approach taken by Omitaomu et al., participants were assigned the task of reading news articles that contained harmful content related to individuals, groups, animals, or objects. Subsequently, they were required to express their reactions in essays and engage in discussions.

For the second iteration of the shared task, a new dataset was introduced. This dataset includes written essays along with associated Batson empathic concern and personal distress scores, as well as the Big Five personality traits (OCEAN) for each reader. Unlike the previous version (Barriere et al., 2023), the new dataset also incorporates conversations between two users who read the same article. Each speech turn in these conversations has been annotated for perceived empathy, emotion polarity, and intensity. Additionally, the dataset provides news articles referenced in the conversations and essays, along with person-level demographic information (age, gender, ethnicity, income, and education level).

The shared task was divided into four subtasks:

- **Task 1: Empathy Prediction in Conversations (CONVD):** Predicting perceived empathy at the dialog level.
- **Task 2: Empathy and Emotion Prediction in Conversation Turns (CONVT):** Predicting perceived empathy, emotion polarity, and intensity at the speech-turn level in a conversation.
- **Task 3: Empathy Prediction (EMP):** Predicting both empathy concern and personal distress at the essay level.
- **Task 4: Personality Prediction (PER):** Predicting the personality traits (openness, conscientiousness, extraversion, agreeableness, and emotional stability) of essay writers based on their essays, dialogs, and the news articles they reacted to.

Our team (Fraunhofer SIT) participated in Tasks 1, 2, and 3. In this paper, we present our solution that combines LLM-driven data augmentation with fuzzy target labels and fine-tuned sentiment transformer models. During the competition, our solution achieved 1st place in Task 1, 8th in Task 2, and 3rd place in Task 3, demonstrating strong performance across empathy classification tasks.

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 435–440
August 15, 2024 ©2024 Association for Computational Linguistics

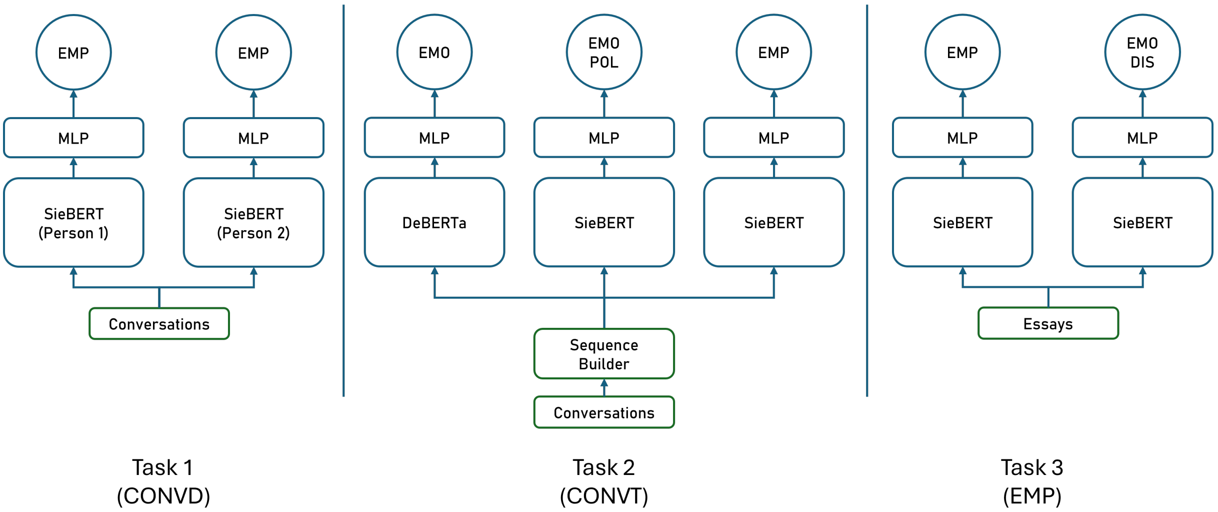

Figure 1: Proposed architectures for each subtask

2 Data Augmentation with Fuzzy Target Labels

Obtaining labeled training data for classification and regression tasks presents challenges. Experts skilled at assigning accurate labels are necessary and, regarding the shared task, enough participants are required willing to engage in discussions and contribute essays about their emotional reactions. Consequently, pre-training language models like BERT (Devlin et al., 2019) on task-specific data is often impractical and can impact fine-tuning applicability. To address this issue, we employed data augmentation in this shared task to generate new samples from the limited existing data. In particular, we focussed on *paraphrasing* and *back-translation* operations.

To maintain independence from external APIs, we performed augmentations using a local instance of *LLama V3 8B-Instruct* [1] (AI@Meta, 2024). For paraphrasing, we used *You are a paraphraser chatbot who just returns the paraphrased input sentences and nothing else!* as a system prompt instructing the model to return paraphrased sentences. For back-translation, *You are a translation chatbot who just returns the translated input sentences into {language} and nothing else! In cases, where translation is not possible, return the original input sentence.* was used to translate the sentences first into *German* and then back to *English*.

Despite being trained on multilingual texts, the translation capabilities of the small model introduce translation errors (Table 1). In this paper, we take advantage of the slight mistranslations to provide new data samples with similar meanings. However, while both operations, paraphrasing and back-translation, rephrase the sentences, either by changing the word order in the sentence and by applying synonym substitution, errors result in minor changes regarding the semantics. As such, it cannot be ensured that the labels associated with the original data sample are still correct. Therefore, we chose to add noise to the labels of the augmented data samples in the range of $[-0.2, 0.2]$ to the labels of Task 3 and noise in the range of $[-0.1, 0.1]$ to the labels of the augmented samples of Task 2. No noise was added to the data of Task 1, as they were provided as hard labels. We chose this particular value for various reasons. First, higher noise led to lower performance on the validation set, whereas too weak noise led to the models overfitting on the content of the data sample text. The results on the Mean-Squared Error (MSE) is displayed in Table 2.

3 System Descriptions

In this section, we present the architectures used to predict the target labels of each respective subtask. An overview is displayed in Figure 1.

3.1 Task 1: Empathy Prediction in Conversations

In a scenario where two people engage in dialogues about a read article, the goal was to predict their empathy levels. These empathy scores were represented as integer labels ranging from 1 to 9. Although classification models are typically used for

[1] LLama 3 8B Instruct

Type	Sentence	Emotion	Emotional Polarity	Empathy
Original (OG)	take care! goodbye	1.3333	0.3333	0.6667
Paraphrase	Farewell! May you be well.	1.2657	0.2986	0.6209
BT: OG - GER	Bleib gesund! Auf Wiedersehen!	1.3278	0.4124	0.6475
BT: OG - GER - ENG	Stay healthy! Goodbye!	1.3302	0.3110	0.7119
Paraphrased BT	Wishing you well! Farewell!	1.2915	0.3305	0.7476

Table 1: Example of LLM-based data augmentation on an utterance of the CONVT dataset

	Empathy	Distress
No Augmentation	2.9321	3.3328
Augmentation	2.9275	2.9101
Fuzzy Augmentation	**2.9193**	**2.3299**

Table 2: Influence of data augmentation on the validation loss (MSE) of the Task 2 (EMP) dataset

	Perceived Empathy	
	r	p
Fraunhofer SIT	**0.193**	0.127
ConText	0.191	0.130
Chinchunmei	0.172	0.173
EmpatheticFIG	0.012	0.923

Table 3: Results on the test set of the CONVD dataset (Task 1). Scores represent Pearson Correlation Coefficients

such predictions, we chose to frame this as a regression problem due to label imbalance in the dataset.

Additionally, we hypothesized a strong correlation between empathy estimation and sentiment. As a result, we conducted experiments by fine-tuning various models:

- **DeBERTa V3 Large:** DeBERTa V3 Large (He et al., 2021)[2] is a model trained on generic data, which improves upon BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) using disentangled attention and enhanced mask decoder and its previous iterations regarding efficiency.

- **SieBERT:** Unlike DeBERTa, SieBERT (Hartmann et al., 2023)[3] (based on a RoBERTa model) was fine-tuned on multiple sentiment estimation datasets. These 15 datasets cover various domains, including reviews and tweets. In experiments, SieBERT significantly outperformed previous related work on a synthetic benchmark dataset.

- **Twitter RoBERTa Base Sentiment:** Furthermore, we experimented with a Twitter RoBERTa Base Sentiment model (Barbieri et al., 2020)[4]. This model, as the name suggests, is built upon the RoBERTa architecture and was specifically trained using Twitter data.

To specify which person's empathy within the conversations should be predicted, a dedicated model was trained. Experiments involving additional tokens to indicate the target label for output did not yield favorable results and were therefore omitted.

To select the best-performing model, we trained multiple instances with different seeds. We used a low learning rate of $1.5e-06$ to align with the fine-tuning purpose. The optimizer employed was *AdamW* (Loshchilov and Hutter, 2019), and the learning rate was dynamically adjusted during training. The model was trained with a batch size of 16 and evaluated on 64 samples per batch. At every 15th step, performance was assessed on the validation set, and early stopping was implemented to mitigate overfitting.

Based on the 500 conversations from the training set and augmented data (including paraphrased, back-translated, and paraphrased back-translated examples), we compared the models' performance. The SieBERT-based model (Avg MSE: 2.205) outperformed the DeBERTa model (Avg MSE: 2.233) and Twitter RoBERTa Sentiment model (Avg MSE: 2.239) on the development set and was chosen for the final submission.

On the test set, the model achieved a Pearson Correlation Coefficient of 0.193, securing the top position in the competition (see Table 3). However, the high p value suggests that the computed r value lacks significance. This highlights the ongoing challenge of accurately estimating empathy at the dialogue level.

[2] DeBERTa V3 Large
[3] SieBERT
[4] Twitter RoBERTa Base Sentiment

	Average	Empathy		Emotion Polarity		Emotion Intensity	
	r	**r**	**p**	**r**	**p**	**r**	**p**
ConText	**0.626**	0.577	0.000	0.679	0.000	**0.622**	0.000
Chinchunmei	0.623	**0.582**	0.000	**0.680**	0.000	0.607	0.000
EmpatheticFIG	0.610	0.559	0.000	0.671	0.000	0.601	0.000
Last_min_submission_team	0.595	0.534	0.000	0.663	0.000	0.589	0.000
hyy3	0.590	0.544	0.000	0.644	0.000	0.581	0.000
Empathify	0.588	0.541	0.000	0.638	0.000	0.584	0.000
empaths	0.477	0.534	0.000	0.422	0.000	0.473	0.000
FraunhoferSIT	-0.007	0.034	0.125	-0.018	0.409	0.032	0.141
Zhenmei	-0.030	-0.027	0.223	-0.020	0.356	-0.043	0.051

Table 4: Results on the test set of the CONV-T dataset (Task 2). Scores represent Pearson Correlation Coefficients

	Average	Empathy		Distress	
	r	**r**	**p**	**r**	**p**
RU	**0.453**	**0.523**	0.000	0.383	0.000
Chinchunmei	0.393	0.474	0.000	0.311	0.004
FraunhoferSIT	0.385	0.375	0.000	**0.395**	0.000
1024m	0.344	0.361	0.001	0.327	0.003
ConText	0.321	0.390	0.000	0.252	0.210
Empathify	0.253	0.290	0.008	0.217	0.049
Daisy	0.213	0.345	0.001	0.082	0.461

Table 5: Results on the test set of the EMP dataset (Task 3). Scores represent Pearson Correlation Coefficients

3.2 Task 2: Empathy and Emotion Prediction in Conversations Turns

The second task involved predicting emotional intensity, polarity, and empathy for each turn in a conversation. The training set comprised $11,166$ turns across 500 conversations provided alongside the shared task.

Recognizing that conversation history significantly influences emotional states at specific turns, the utterances were not classified individually. Instead, they were considered along with previous utterances within the conversation. To focus on utterances impacting the current emotional state, a context window of size 5 was used to create sequences of turns.

The models were trained similarly to those in Task 1, with the addition of introducing noise into augmented samples in the range of $[-0.1, 0.1]$. In experiments, the DeBERTa model excelled in classifying emotion on the validation set (Pearson Correlation Coefficient: 0.313), while fine-tuned SieBERT models performed best for emotional polarity (Pearson Correlation Coefficient: 0.3057) and empathy classification (Pearson Correlation Coefficient: 0.282).

During the evaluation on the test set, it was revealed, that the model was unable to provide correct classifications (Table 4). One reason for this might be, that the model was unable to learn significant information based on the short context windows, as indicated by the low validation scores. Future directions may incorporate combining larger context windows as well as the combination of information on turn and dialog level.

3.3 Task 3: Empathy Prediction

Our participation in the last task focused on estimating empathy and emotional distress related to a read article. Participants wrote essays expressing their feelings after reading the article. The training set included 1000 essays, while 66 essays were reserved for the development set.

During training, we fed the essays and their augmentations into individual models. Unlike the previous task, we introduced higher label noise ($[-0.2, 0.2]$). In addition, the batch size during training was raised to 48.

On the development set, the SieBERT models performed the best, with the Pearson Correlation Coefficient for the empathy being 0.6871 and for the emotional distress 0.684. In comparison, the best scores obtained by the DeBERTa model for the empathy estimation was 0.525 and for the emotional distress 0.5602. Using a sentiment transformer pretrained on Twitter data did improve the performance. The best performing Twitter RoBERTa Base Sentiment classifier achieved a Person Correlation Coefficient score of 0.6316 for the empathy estimations and a score of 0.6517 for the emotional distress detection. This shows not only the effectiveness of the sentiment transformers in solving these tasks, but also that the models perform better when trained on sentiment datasets consisting of data from different domains.

In the competition, our proposed system ranked third overall. While it excelled in predicting emotional distress compared to other systems, it fell short in classifying empathy, where it ranked fourth.

4 Conclusion and Future Work

In this paper, we presented the solutions developed by our team Fraunhofer SIT for the 2024 shared task of the *Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Our experiments revealed that models fine-tuned for sentiment estimation tasks often outperformed larger language models, such as DeBERTa, which were trained on more generic data. Data augmentation improved classification accuracy, and introducing noisy labels further refined performance. While our solutions achieved 1st place in Task 1, 8th in Task 2, and 3rd place in Task 3, the Pearson Correlation Coefficients indicate the need for additional research to achieve more stable results.

5 Limitations

Experiments have shown that solving the empathy and emotion estimation tasks poses various challenges. In the particular case of Track 1 (CONV-dialog), the performance of proposed the model according to the Pearson Correlation Coefficients is low despite the first place in the competition. One reason for this is that many of the models used were unable to predict meaningful labels during training. Instead, target labels that deviated from the mean were often incorrectly predicted. The transition from a regression to a classification problem did not solve the problem. This indicates that the imbalance of the labels often has a significant impact on the performance of the models.

Furthermore, the performance of the models depends on the seeds used. Training the models with different seeds leads to different results. However, taking advantage of data augmentation always led to an increase in performance.

Although the fine-tuned sentiment transformers based on SieBERT often performed best, the Twitter RoBERTa base sentiment models did not. This suggests that texts in tweets are stylistically too different from those in essays and even dialogs. Therefore, it is recommended to use sentiment transformers trained on general texts or texts from different data sources and domains.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of “ATHENE – CRISIS” and "ATHENE CHARM".

References

AI@Meta. 2024. Llama 3 model card.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. 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*.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

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. *Preprint*, arXiv:2205.12698.