Team_Hawk at WASSA 2023 Empathy, Emotion, and Personality Shared Task: Multi-tasking Multi-encoder based transformers for Empathy and Emotion Prediction in Conversations

Addepalli Sai Srinivas, Nabarun Barua, Santanu Pal
Wipro AI Research (Lab45), India
{addepalli.srinivas, nabarun.barua, santanu.pal2}@wipro.com

Abstract

In this paper, we present Team Hawk's participation in Track 1 of the WASSA 2023 shared task. The objective of the task is to understand the empathy that emerges between individuals during their conversations. In our study, we developed a multi-tasking framework that is capable of automatically assessing empathy, intensity of emotion, and polarity of emotion within participants' conversations. Our proposed core model extends the transformer architecture, utilizing two separate RoBERTa-based encoders to encode both the articles and conversations. Subsequently, a sequence of self-attention, position-wise feed-forward, and dense layers are employed to predict the regression scores for the three sub-tasks: empathy, intensity of emotion, and polarity of emotion. Our best model achieved average Pearson's correlation of 0.7710 (Empathy: 0.7843, Emotion Polarity: 0.7917, Emotion Intensity: 0.7381) on the released development set and 0.7250 (Empathy: 0.8090, Emotion Polarity: 0.7010, Emotion Intensity: 0.6650) on the released test set. These results earned us the 3^{rd} position in the test set evaluation phase of Track 1.

1 Introduction

Empathy involves understanding and sharing others' feelings. In conversation, empathy is demonstrated through active listening, acknowledging emotions, and providing supportive responses. Emotion polarity refers to the positive or negative nature of expressed emotions, while emotion intensity relates to the strength of those emotions.

Computing empathy is an emergent paradigm and become an important component in conversational AI (Mazaré et al., 2018; Roller et al., 2021). Empathy is critical for clinical applications such as automated behavioral therapy (Fitzpatrick et al., 2017). Implementing complex emotional-motivational states and effectively responding in an empathetic manner remains a significant challenge in human-machine interaction.

The dataset shared by the organizers comprises news stories and corresponding brief essays written by participants during conversation sessions. Participants engage in dialogues, assessing each other's conversation turns for empathy, emotion intensity, and polarity. A third-party annotator confirms the emotional dimensions of empathy, intensity, and polarity at the end of the session (Buechel et al., 2018; Sharma et al., 2020; Barriere et al., 2022).

The WASSA 2023 shared task (Barriere et al., 2023) consists of five different tracks: Track 1: Empathy and Emotion Prediction in Conversations (CONV), Track 2: Empathy Prediction (EMP), Track 3: Emotion Classification (EMO), Track 4: Personality Prediction (PER), and Track 5: Interpersonal Reactivity Index Prediction (IRI).

In this paper, we present our submission for Track 1. The provided dataset contains conversations between two participants, along with scores for empathy, emotion polarity, and emotion intensity. The task revolves around two participants who read a news article and initiate a conversation based on that particular news. They are then required to assess each other's empathy, intensity of emotion, and polarity of emotion based on their discussion.

The objective of the task is to develop a system capable of automatically assessing empathy, intensity of emotion, and polarity of emotion within participants' conversations.

Our model consists of two RoBERTa-based encoders to encode article and conversation followed by a 3-layered transformer encoder. The representation is then passed to a sequence of layers which provides regression output of three tasks (cf. Track 1) – Empathy, Emotion Polarity, and Emotion Intensity with a multi-tasking framework (MLT). In Track 1 participation, our best model achieved average Pearson's correlation of 0.771 on the released development set and 0.725 on the released test set. The key findings of this research can be summarized as follows: (i) residual skip connections are

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 542–547
July 14, 2023 ©2023 Association for Computational Linguistics

effective in enhancing the conversation encoder, (ii) including the previous dialogues of both participants in a conversation along with the current dialogue helps in preserving the context of the conversation within the dialogue encoder for a particular session, and (iii) Token interactions between articles and conversations utilizing multi-head self-attention yield significant and informative results.

The rest of this paper is organized as follows. We introduce related work in §2. We discuss problem statement in §3. The proposed model is described in §4 and experiment and result in § 5. Finally, we conclude our paper in §6.

2 Related Work

Recently, transformer based (Vaswani et al., 2017) pre-trained models such as BERT (Devlin et al., 2019), OpenAI GPT (Radford et al., 2018), RoBERTa (Liu et al., 2019) etc. have been shown superior performance in various downstream tasks, including text classification task (Sun et al., 2020; Luo and Wang, 2019; Singh et al., 2021), generation task such as question answering (Garg et al., 2020) and many more. Recent works have shown that using such pre-trained methods can achieve state-of-the-art performance. Towards that end, Sharma et al. (2020) investigated a multi-task RoBERTa-based bi-encoder paradigm for comprehending empathy in text-based health support, Zhou and Jurgens (2020) investigated the link between distress, condolence, and empathy in online support groups using nested regression models. Many research (Abdul-Mageed and Ungar, 2017; Nozza et al., 2017) have given various strategies for emotion recognition. The effectiveness of using transformer encoders for emotion detection was investigated by Adoma et al. (2020). Ghosh et al. (2022) proposed a multi-task deep learning methods to address Empathy Detection, Emotion Classification and Personality Detection. Inspired from (Sharma et al., 2020) and multi-encoder based architectures (Pal et al., 2018, 2019, 2020), we propose a multi-encoder based architecture followed by MLP and a linear layer output layer. Our core architecture is similar to Pal et al. (2018), the difference is we use two RoBERTa (Liu et al., 2019) encoders for inputs.

3 Problem Statement

Our model is based on a multi-task learning based framework (MTL) to force the model to consider three different objectives i.e. three emotional dimensions of empathy (y_1), intensity (y_2), and polarity (y_3). Given a set of conversation for a single session $\mathbf{c}$ for a corresponding article $\mathbf{a_c}$, the output probability $\mathbf{y} \in y_1, y_2, y_3$ in the model setting is calculated as in Equation 1.

$$p(\mathbf{y}) = p(\mathbf{y}|\mathbf{c}, \mathbf{a_c}) \tag{1}$$

The model acts as a regressor, the output head provides regression results. The network consists of a two-layered multi-layered perceptron[1] (MLP) with ReLU activation between layers and output head with a 'linear' activation. Given the output of the regression head, the loss can be calculated as:

$$\mathcal{L}_t = \sum_{(\mathbf{a_c}, \mathbf{c}, \mathbf{y}) \in \mathcal{D}} -logp(\mathbf{y}|\mathbf{c}, \mathbf{a_c}; \theta) \tag{2}$$

Our model is trained for three emotional dimensions of empathy (y_1), intensity (y_2), and polarity (y_3) with corresponding losses $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ respectively, in an end-to-end fashion that jointly optimizes the loss as in Eq. 3, where α, β, and γ are learnable parameters.

$$\mathcal{L}_{overall} = \alpha * \mathcal{L}_1 + \beta * \mathcal{L}_2 + \gamma * \mathcal{L}_3 \tag{3}$$

4 Model Architecture

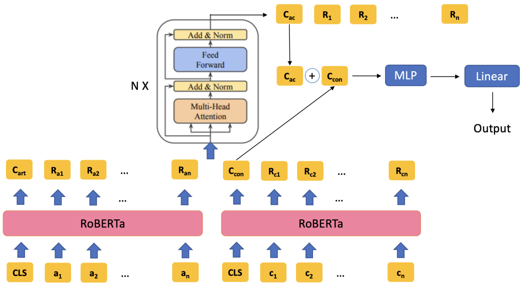

Figure 1: Our proposed model architecture

Our proposed model consists of two Roberta encoders[2] – article encoder (enc_{art}) and conversation encoder (enc_{conv}) based on the inputs from news articles and conversations. In the proposed model, the last hidden state of both encoder representations are merged and passed through a 3-layered transformer encoder (cf. Figure 1). A residual connection is employed between the CLS pooling of the conversation encoder and the output of the 3-layered transformer encoder. The combined CLS

[1] Our MLP layer consists of two fully connected feed forward layers and a ReLU activation between them.

[2] For our submission, we use Roberta Large: https://huggingface.co/roberta-large

pooling representation is then inputted into MLPs followed by a linear layer. The linear layer generates regression outputs for the three tasks: Empathy, Emotion Polarity, and Emotion Intensity. This model utilizes a multi-tasking framework (MLT) that jointly optimizes each individual loss function (see §3) in an end-to-end manner.

5 Experiments and Results

In this section, we provide a summary of how our systems were trained, tuned, and combined to create the Team Hawk submissions for Track 1 of the WASSA 2023 Shared Task. We evaluated our system using Pearson's correlation, which measures the relationship between the regression outputs of our model and the gold standards. All benchmark evaluation scores are reported based on the development set released by the organizers.

5.1 Dataset

The track 1 dataset (Barriere et al., 2023; Omitaomu et al., 2022) used in this research consists of conversations between two individuals within a given session. The dataset includes several columns, such as conversation-id, turn-id, conversation text, emotional polarity, emotion, empathy, speaker-number, article-id, speaker-id, essay-id, and more. Additionally, an article dataset is also utilized, which contains article-id and article-text.

The overall training data comprises 8,778 labeled data points. To conduct the experiments, the released training set is split into a validation set, consisting of 1,756 data points, and a train set, containing the remaining 7,022 data points.

For evaluation purposes, the released development data consisting of 2,400 data points is used as the in-hand test set.

5.2 Data Pre-processing

To prepare inputs to our model we performed data preprocessing. The enc_{art} takes input as articles with <a_id:article_id>. To the enc_{conv}, we provide the contextual conversations (C) along with Conversation ID, Turn ID, Article ID & Speaker ID represented as <c_id:conversation_id>, <tid:turn_id>, <a_id:article_id> & <s_id:speaker_id> respectively for each conversation (c). For i^{th} training instance for a particular session the contextual conversion means $C = concat(c_{<i}, c_i)$.

5.2.1 Baseline Model

Our baseline model is similar to our proposed model however the model does not have transformer encoder (cf. Figure 2).

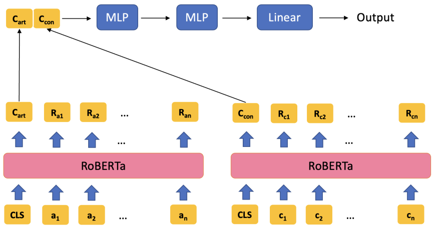

Figure 2: Baseline Model

The baseline model demonstrated an average Pearson correlation of 0.623 on the validation set. However, on the training set, it achieved an average Pearson correlation of approximately 0.977. This significant difference between the performance on the training set and the validation set indicates clear overfitting, as the model has excessively adapted to the training data.

5.2.2 Baseline with residual connection

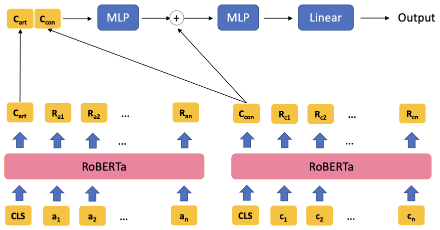

Figure 3: Residual Model

In this experiment, we add a residual connection between MLP representation of combined encoders (concatenation of enc_{art} and enc_{conv}) and the conversation encoder, enc_{conv} (cf. Figure 3). In our model, the combination of both encoders through CLS pooling raises concerns about the possibility of information loss that has been learned by the enc_{conv}. To address this issue, we utilize skip connections, which not only help alleviate the vanishing gradient problem but also improve training efficiency. However, the key question is why we use skip connections in our model.

The use of skip connections ensures that the gradient flows equally to both the article and dialogue encoders. While the article remains constant, empathy, emotion polarity, and emotion intensity are primarily driven by the conversation. Therefore, it is crucial to emphasize the importance of the conversation in our model to provide a diverse range of contexts for the model to learn from.

Furthermore, introducing residual connections offers another avenue for improvement in the output head. With the presence of multiple linear layers, there is a risk of overfitting. The inclusion of residual connections helps mitigate this concern and enhances the model's performance by allowing the flow of information from earlier layers to later layers, enabling better representation learning.

After combining the representations from the two encoders, we proceed with the following steps:

Minimizing the output feature: Instead of directly passing the input through multiple linear layers, we utilize two Multi-Layer Perceptrons (MLPs). Each MLP consists of two linear layers with ReLU activation. The first MLP reduces the output feature dimension from $\mathbf{R}^{1536}$ to $\mathbf{R}^{768}$ (specific to the RoBERTa base). This reduced feature is then passed through Layer Normalization to enhance training stability and speed.

Further feature reduction: The output from Layer Normalization is fed into the second MLP, which reduces the feature dimension from $\mathbf{R}^{768}$ to $\mathbf{R}^{32}$. Layer normalization is once again applied to the output to maintain stability.

Final linear layer: The output from the second MLP is then passed through a linear layer, which further reduces the feature size from $\mathbf{R}^{32}$ to $\mathbf{R}^{1}$.

These steps enhance the model's to an improved output for regression tasks. As a result, we achieved a significant increase in the Pearson correlation, improving it from 0.623 to 0.724.

5.2.3 The proposed model

Here, we introduced a 3-layered transformer encoder (cf. §4). The main difference from the previous experiment is that instead of passing the concatenation of the CLS pooling of enc_{art} and enc_{conv} to the MLP, we pass the combined last hidden state ($\mathbf{R}^{512\times1536}$) of both encoders (for RoBERTa base, the representation of $enc_{art} \in \mathbf{R}^{512\times768}$ and $enc_{conv} \in \mathbf{R}^{512\times768}$) to 3 layered transformer encoder (cf. Figure 1). Similar to the previous experiment, we include a residual connection between the CLS Pooling of the 3-layered transformer encoder ($\mathbf{R}^{768}$) and the CLS Pooling of the conversation encoder ($\mathbf{R}^{768}$). Finally, the combined representation ($\mathbf{R}^{1536}$) is then inputted to an MLP, followed by a Linear Layer that generates the regression output (see steps in §5.2.2).

5.2.4 Result and Discussions

Our submission results are shown in Table 1. As discussed in Section 5.2.1 and 5.2.2, we can see our proposed approach provides best Pearson's scores on development data compared to both baseline and baseline with residuals. We also report testset score evaluated by the organizers. Out submission ranked 3^{rd} in Track 1. In this submission, we use RoBERTa-large for both encoders i.e., enc_{art} and enc_{conv} (cf. Figure 1), with maximum token length 512. Other hyper-parameters include batch size = 6, learning rate = $5e^{-5}$, number of epochs = 30. For our custom transformer encoder, we set number of layers = 3, embedding size = 1024, and head size = 8. All models are trained with mean-squared error loss criteria and optimized with default configuration of Adam optimizer.

Data	Emp	Emo-Pol	Emo-Int	Avg
Dev	0.7843	0.7917	0.7381	0.7710
Test	0.8090	0.7010	0.6650	0.7250

Table 1: Performance of our submission based on Pearson's score on *Development* & *Test sets*. Here, we use RoBERTa-large for enc_{art} and enc_{conv}, with maximum token length 512. Emp: Empathy Emo-Pol: Emotion polarity, Emo-Int: Emotion Intensity, Avg: Average.

6 Conclusion and Future Work

In this paper, we introduced our methodologies for investigating the emergence of empathy during conversations. The task is introduced as part of the WASSA 2023 shared task track 1, which focuses on Empathy and Emotion Prediction in Conversations (CONV). We developed a multi-tasking framework that leverages a core multi-encoder based architecture. This framework enables automatic assessment of empathy, intensity of emotion, and polarity of emotion in participants' conversations. Our systems achieved average Pearson's scores of 0.7710 on the released development set and 0.7250 on the released test set. Our submission ranked 3^{rd} in the shared task. Due to time constraints, we are unable to conduct an exhaustive number of experiments with various architecture variations before reaching a final conclusion. However, for future research, we will conduct a comprehensive analysis involving architecture variations and data preprocessing methods. Additionally, we plan to investigate the influence of other features such as gender, age, etc., on the model's decision-making process.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Acheampong Francisca Adoma, Nunoo-Mensah Henry, Wenyu Chen, and Niyongabo Rubungo Andre. 2020. Recognizing emotions from texts using a bert-based approach. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 62–66.

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*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

KK Fitzpatrick, A Darcy, and M Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. In *JMIR Ment Health.*

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: transfer and adapt pre-trained transformer models for answer sentence selection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7780–7788. AAAI Press.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Linkai Luo and Yue Wang. 2019. Emotionx-hsu: Adopting pre-trained bert for emotion classification.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. 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.

Santanu Pal, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2018. A transformer-based multi-source automatic post-editing system. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 827–835.

Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2019. Usaar-dfki–the transference architecture for english–german automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 124–131.

Santanu Pal, Hongfei Xu, Nico Herbig, Sudip Kumar Naskar, Antonio Krüger, and Josef van Genabith. 2020. The transference architecture for automatic post-editing. In *28th International Conference on Computational Linguistics*, 10.18653/v1/2020.coling-main.524, pages 5963–5974. https://www.aclweb.org/anthology/2020.coling-main.524.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

G. Singh, D. Brahma, P. Rai, and A. Modi. 2021. Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Los Alamitos, CA, USA. IEEE Computer Society.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, Online. Association for Computational Linguistics.