# EmoNet-XAI: An Explainable First-Place Approach to Transformer-based Emotion Classification

Md. Abdur Rahman
*Department of Computer Science and Engineering*
*Southeast University, Bangladesh*
Email: 2021200000025@seu.edu.bd

***Abstract*—This report presents a detailed methodology for EmoNet-XAI, the first-place solution developed by SmolLab_SEU for the "Fragments of Feeling" Kaggle competition. The framework is built upon the RoBERTa-Large transformer model and introduces a synergistic combination of advanced fine-tuning techniques, a robust evaluation strategy, and a commitment to model transparency. Key innovations include the implementation of Layer-wise Learning Rate Decay (LLRD) for stable and effective optimization, a 5-fold stratified cross-validation ensemble to enhance generalization and reduce variance, and the integration of eXplainable AI (XAI) using SHAP for model interpretability. This meticulously engineered pipeline achieved a final Macro F1-Score of 0.6306 on the private leaderboard, securing the top rank. All code is available on GitHub: https://github.com/borhanitrash/EmoNet-XAI.**

***Index Terms*—Emotion Classification, Transformer Models, Layer-wise Learning Rate Decay, Explainable AI (XAI), Ensemble Method, SHAP**

## I. Introduction

The task of classifying human emotions from short text requires a deep understanding of syntax, semantics, and contextual nuance. Pre-trained Language Models (PLMs) such as BERT and its variants have become the standard for such Natural Language Understanding (NLU) tasks due to their ability to learn rich, contextualized representations of text [1]. However, unlocking their full potential requires more than simple fine-tuning.

Our proposed solution, EmoNet-XAI, is a comprehensive framework designed to maximize predictive performance while ensuring the model's decisions are transparent and interpretable. We hypothesize that a combination of a powerful base model, a sophisticated fine-tuning strategy that respects the hierarchical nature of PLMs, and a robust ensemble method is key to achieving state-of-the-art results. Furthermore, we posit that incorporating an explainability layer is not merely an add-on but a crucial component for building trustworthy AI systems. This report details the architecture, training procedures, and novel components that led to our winning performance.

## II. Methodology

The EmoNet-XAI framework follows a systematic pipeline, from data preparation to final prediction, as illustrated in Fig. 1.

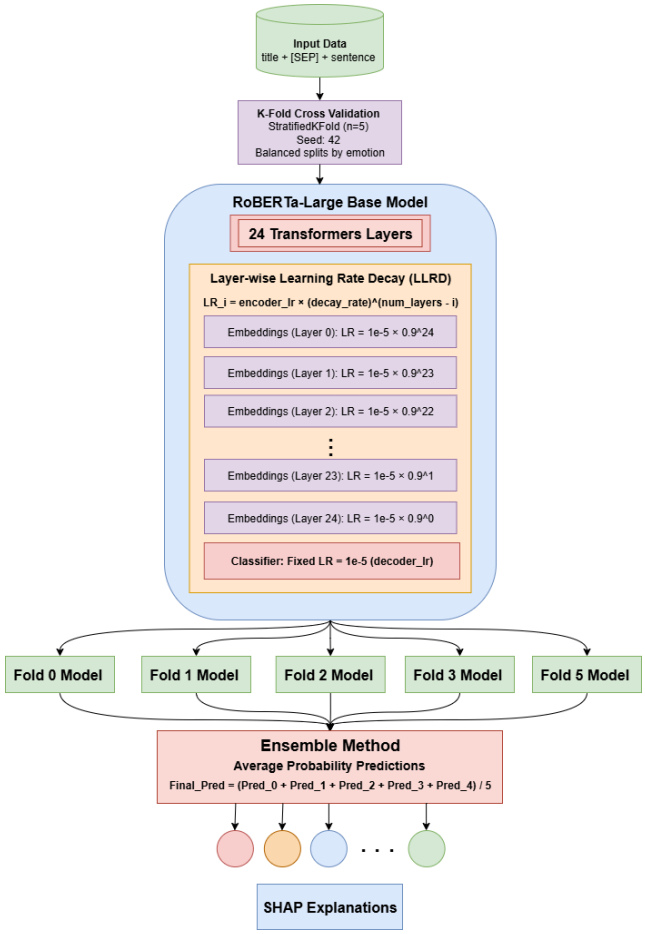


Fig. 1. The Proposed EmoNet-XAI Framework

### A. Base Model Selection

The foundation of our framework is the RoBERTa-Large model [2]. We selected RoBERTa over the original BERT for several reasons: it was pre-trained on a significantly larger corpus of text, used a dynamic masking strategy, and removed the Next Sentence Prediction (NSP) objective, all of which have been shown to improve performance on downstream tasks. The 'Large' variant, with its 24 transformer layers, 1024 hidden dimensions, and ˜355M parameters, provides the necessary model capacity to capture the intricate patterns present in the emotion-laden text.

### B. Data Preprocessing and Feature Engineering

The input data consisted of a title and a sentence. Recognizing that the title often provides crucial context for interpreting the emotion of the sentence, we engineered a single input sequence by concatenating them. The fields were separated by the model's special [SEP] token: [title] [SEP] [sentence]. This structure allows the model's self-attention mechanism to explicitly learn the relationship between the title and the sentence. The combined text was then tokenized using the AutoTokenizer associated with FacebookAI/roberta-large. Sequences were padded or truncated to a maximum length of 256 tokens, a length determined to be sufficient to cover the vast majority of samples without incurring excessive computational cost.

### C. Training Strategy

Our training strategy was designed for robustness, stability, and optimal convergence.

To build a model that generalizes well to unseen data and to obtain a reliable estimate of its performance, we employed a 5-fold Stratified Cross-Validation scheme. The dataset was split into five folds, ensuring that the proportion of samples for each emotion class was maintained across all folds. This is particularly important for mitigating biases that could arise from class imbalances. We trained five independent RoBERTa-Large models from scratch, with each model using a different fold as its validation set and the remaining four folds as its training set.

A central innovation of our method is the application of Layer-wise Learning Rate Decay (LLRD), a form of discriminative fine-tuning [3]. The core principle is that different layers of a PLM learn different types of information; lower layers capture general linguistic features, while higher layers learn more task-specific representations. Fine-tuning all layers with the same learning rate risks either "catastrophic forgetting" in the lower layers or insufficient adaptation in the higher layers. LLRD addresses this by applying progressively smaller learning rates to lower layers. We implemented a multiplicative decay schedule where the learning rate for encoder layer $i$ ($\text{LR}_\text{i}$) is defined as:

$$\text{LR}_i = L_{\text{base}} \times \alpha^{(L_{\text{total}} - i)},$$

where $L_{\text{base}}$ is the base learning rate for the top encoder layer (1e-5), $\alpha$ is the decay rate (0.9), $L_{\text{total}}$ is the total number of transformer layers (24), and $i$ is the layer number. The final classification head (decoder) was assigned a separate, fixed learning rate of 1e-5. This strategy promotes stable fine-tuning, preserving valuable pre-trained knowledge while allowing task-specific layers to adapt more aggressively.

The optimization process was governed by the AdamW optimizer [4], chosen for its effective decoupling of weight decay from the gradient update, which leads to better regularization. To further enhance generalization and prevent the model from becoming overconfident, we employed Label Smoothing with a factor of 0.1 [5]. This technique discourages the model from producing extreme logit values. The training process was also stabilized by using gradient accumulation with 2 steps. This allowed us to achieve an effective batch size of 16 using a per-device batch size of 8, which helped manage GPU memory constraints while maintaining training stability.

TABLE I
TRAINING HYPERPARAMETERS FOR EMONET-XAI

| Hyperparameter | Value |
|---|---|
| Seed | 42 |
| Folds | 5 |
| Epochs | 5 |
| Max Sequence Length | 256 |
| Per-Device Batch Size | 8 |
| Gradient Accumulation | 2 |
| **Effective Batch Size** | **16** |
| Optimizer | AdamW |
| Base Encoder LR | 1e-5 |
| Classifier LR | 1e-5 |
| LLRD Decay Rate | 0.9 |
| Weight Decay | 0.01 |
| Label Smoothing Factor | 0.1 |
| Mixed Precision (FP16) | Enabled |

### D. Hyperparameter Configuration

All models were trained with a consistent set of hyperparameters, summarized in Table I.

## III. EVALUATION AND INFERENCE

### A. Evaluation Metric

The official competition metric was the *Macro F1-Score*. This metric calculates the F1-score for each class independently and then takes the unweighted average. It is an ideal metric for multi-class classification, as it treats all classes as equally important, preventing a model's performance on a majority class from dominating the overall score.

### B. Ensemble Method

For final inference on the unseen test set, we ensembled the five models trained during cross-validation. An ensemble approach typically yields more robust and accurate predictions by reducing the variance of a single model's predictions. The inference process began with preprocessing the test data identically to the training data. For each sample in the test set, predictions in the form of raw logits were obtained from each of the five trained models. These logits were then averaged element-wise to produce a single, ensembled logit vector. Finally, the predicted emotion class was determined by applying the *argmax* function to this ensembled vector, a simple yet highly effective method for smoothing out individual model biases.

### C. Explainable AI (XAI) with SHAP

A defining feature of our framework is its integrated explainability component. To move beyond a "black box" solution, we employed SHAP (SHapley Additive exPlanations) [6]. SHAP is a game-theoretic approach that computes the contribution of each feature in our case, each token to a model's prediction. This capability provides invaluable,

human-interpretable insights into the model's reasoning, allowing for qualitative analysis and building trust in its outputs. The power of this approach is best illustrated through specific case studies.

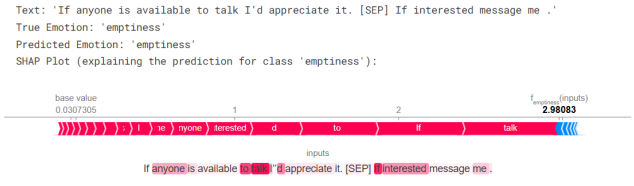


Fig. 2. SHAP analysis for a text expressing a need for connection. Red tokens increase the probability of the 'emptiness' class.

As shown in Figure 2, for the input *"If anyone is available to talk I'd appreciate it. [SEP] If interested message me."*, the model correctly predicts 'emptiness'. The SHAP analysis reveals that the model's decision is not based on overtly negative keywords. Instead, it correctly identifies the implicit sentiment of loneliness. Tokens such as "anyone", "available to talk", and "interested message me" are highlighted in red, indicating they strongly pushed the prediction towards 'emptiness'.

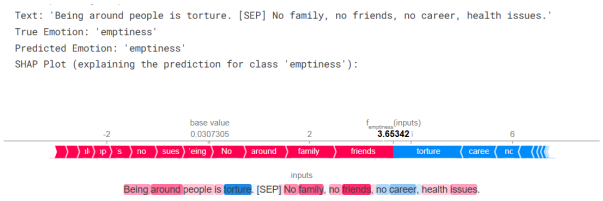


Fig. 3. SHAP analysis for a text describing social isolation. The model weighs conflicting emotional signals.

Figure 3 presents a more complex case of emotional reasoning. For the input *"Being around people is torture. [SEP] No family, no friends, no career, health issues."*, the model again correctly predicts 'emptiness'. The analysis shows that the primary drivers for this prediction are the phrases "No family" and "no friends", which directly signify social isolation. Tokens highlighted in blue pushed the prediction away from 'emptiness'. The highly negative word "torture" is surprisingly a counter-indicator, suggesting the model associates it more with other emotions such as sadness or anger. This demonstrates the model's advanced reasoning capabilities.

## IV. Conclusion

The first-place performance of EmoNet-XAI is a testament to a principled and multi-faceted approach. The success of the framework can be attributed to the confluence of several key factors. Its foundation rests on a sophisticated fine-tuning strategy, where the use of Layer-wise Learning Rate Decay (LLRD) allowed for a more stable and effective adaptation of the powerful RoBERTa-Large model. This was complemented by a focus on robust generalization, achieved through a 5-fold stratified ensemble methodology that ensured the final predictions were resilient to data idiosyncrasies. Finally, a commitment to transparency was demonstrated by the integration of SHAP, which provided a crucial layer of interpretability and transformed the model from a "black box" into an explainable system. By meticulously engineering each component, EmoNet-XAI demonstrates that state-of-the-art performance is achieved through a synergy of model capacity, advanced training techniques, and a commitment to model transparency.

## References


[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339. [Online]. Available: https://aclanthology.org/P18-1031/

[4] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.