RoBERTuna: Emotion Detection in Text with Optuna Optimization and LLM-Assisted Evaluation

Kyuri Im (5179237) Lyoungah Kim (5192778)
Technische Universität Dresden
kyuri.im@mailbox.tu-dresden.de lyoungah.kim@mailbox.tu-dresden.de

February 10, 2025

Abstract

Emotion detection from textual data is a critical task in natural language processing (NLP), with applications spanning artificial intelligence, human-computer interaction, psychology, marketing. As part of the SemEval 2025 competition, this paper presents RoBERTuna, a multilabel emotion detection model based on RoBERTa. A dynamic thresholding technique is introduced to adaptively determine the decision boundary for emotion classification, addressing the variability in emotion intensity between inputs. RoBERTuna is integrated with the Optuna, which is an advanced hyperparameter optimization framework, to derive the best parameters that maximize the model performance. RoBERTuna shows satisfactory performance within a small number of epochs and thus provides high computational efficiency and time efficiency. The code and evaluation results are available at `https://github.com/kim33/LLMProject_SemEval`

1 Introduction

Emotions have been categorized using psychological models such as Plutchik's eight primary emotions [6] or Ekman's six basic emotions [3]. . Emotion detection has gained significant attention due to its applications in artificial intelligence, human-computer interaction, psychology, and marketing. However, emotion detection is yet another challenging task in natural language processing (NLP), involving the identification of discrete emotions based on word choice, nuance, and overall context. Words and sentences can carry multiple meanings and perspectives, making precise classification complex. Additionally, many systems classify emotions using a fixed set of labels, whereas human emotions are far more nuanced and may overlap. This variability is further influenced by individual background knowledge and personality [1]. To address these challenges, modern approaches leverage deep learning models such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers like BERT and GPT, which incorporate contextual understanding and improve classification accuracy. Researchers have also explored multi-label classification techniques, allowing multiple emotions to be assigned to a single text to better capture its complexity [10]. In this paper, as part of SemEval-2025 emotion detection task, we adopt a multi-label classification approach with k-fold cross-valiadation and the Optuna framework to maximize the accuracy of emotion detection.

2 Related Work

2.1 RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) is an enhanced version of BERT, developed by Facebook AI.[8] On top of the advantages of transformer architecture, RoBERTa

understands the long-range dependencies in text so that it captures how words are related to each other. Moreover, because RoBERTa is pre-trained with a vast amount of text data with a larger batch size and dynamic masking strategy during the training, it allows to be specialized in classification tasks, and also the model learns the context of the language, which is critical for emotion detection in text. RoBERTa also utilizes the [CLS] token, which is added at every start of the input sequence. [CLS] token lets RoBERTa be an optimal model for classification tasks by aggregating the information from all parts of the input through self-attention. During the training, the final hidden state corresponding to this token is treated as a summary representation of the entire sequence. When fine-tuning RoBERTa for emotion detection, a classification head is added on top of the [CLS] token representation and learns to map the aggregated features to specific emotion categories.

2.2 Optuna

Optuna is an open-source hyperparameter optimization framework designed to automate and efficiently search for the best hyperparameters in machine learning models. With ease of implementation with PyTorch, Keras, TensorFlow, and Scikit-Learn, Optuna provides a flexible and dynamic approach to tuning models using pruning and adaptive searching techniques. [2] With the sampling strategy, Optuna uses techniques like Tree-structured Parzen Estimator (TPE) for efficient exploration, and it automatically stops unpromising trials and prunes search spaces through the pruning strategy, saving computational resources efficiently. Because of its high flexibility to other libraries and high efficiency in search algorithm, Optuna has been used in emotion detection with LLM, such as EmoBERTa[7], Categorical Emotion Detection combining BERT and ChatGPT models[5], SAMSEMO [4], physiological-emotion recognition model[12], and many more.

3 Methodology

In this section, we provide an overview of the SemEval training dataset and outline our approach to leveraging RoBERTa and Optuna. By integrating the strengths of these tools, we aim to optimize performance in the emotion detection task.

3.1 Dataset

The train dataset is provided by SemEval-2025. From the available multilingual data, we specifically focus on the English text subset. This dataset contains 2,769 text entries, each classified into one or more emotion categories: anger, fear, joy, sadness, and surprise. It is important to note that some texts are not associated with any of these emotions. A comprehensive paper with details about the data collection process, annotation methodology, and baseline experiments for the training dataset will be published by the SemEval organization.

3.2 Training

K-Fold Cross Validation: Given the limited size of our training dataset, K-Fold Cross Validation is applied to optimize the use of available data with 5 folds and 10 folds to evaluate model performance under different partitioning schemes. To mitigate the risk of overfitting, an EarlyStoppingCallback is set with patience of 2, that halts the training process if the validation loss fails to improve over two consecutive epochs.
Optuna: The Optuna framework is applied to automate the hyperparameter tuning process of the model. The objective function defined in Optuna explores key hyperparameters such as learning rate, dropout rate, batch size, and the number of epochs within user-defined limits. The table 1 shows the range of hyperparameters implemented in RoBERTuna. It returns the

optimal set of hyperparameters that maximize the F1 score, ensuring the best model performance.

Learning Rate	Drop-out Rate	Batch Size	Epochs
5e-5, 1e-4	0.1, 0.2	8, 16	3, 6

Table 1: Optuna Parameter Option

Dynamic Threshold: A dynamic threshold is used to determine the final predictions, based on the average probability of emotions for each text. This dynamic approach allows the model to adapt to variations in data distribution, enhancing its robustness across different contexts. Additionally, dynamic thresholding accounts for contextual elements such as tone and word frequency, enabling greater sensitivity to subtle or weaker emotions. The model assigns a value of '1' to an emotion if its probability exceeds the dynamic threshold; otherwise, it assigns '0'.
Loss Function and Optimizer: To optimize its ability to capture nuanced emotional expressions in text, Cross-entropy loss and Adam Optimizer are integrated into RoBERTa. The Cross Entropy Loss function quantifies the difference between predicted and actual emotion labels, ensuring robust classification performance. The Adam optimizer is utilized to adjust model weights efficiently, leveraging adaptive learning rates to enhance convergence stability.

4 Evaluation

4.1 F1 Score

The F1 score provides a harmonic mean of precision and recall, where precision is the ratio of true positive predictions to the total predicted positives and recall is the ratio of the true positive predictions to the total actual positives. Both k=5 and k=10 reach to the best performance within 6 epochs. k=10 shows slightly higher results in overall metrics, but the difference compared to k=5 is not significant. Considering the computational and time efficiency, we conclude that k=5 shows a more satisfying result.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

	Epochs	F1 Score	Accuracy	Loss	Precision	Recall
k=5	6	0.95114	0.95679	0.05381	0.94097	0.96434
k=10	6	0.95259	0.95809	0.04660	0.94251	0.96560

Table 2: F1 Score of RoBERTuna with k=5 and k=10

4.2 Confusion Matrix

To assess the classification performance of the model, we use the Confusion Matrix, which provides a detailed breakdown of the model's predictions. The matrix shows the true positive (TP), true negative (TN), false positive (FP), and false negative (FN), and therefore provides not only the overall accuracy of the model but also its performance across different classes, highlighting areas where the model may be misclassifying instances.

Overall, both Figure 1 and Figure 2 show diagonal shades with true positives among the predicted emotions compared to the expected values. The model shows the highest number of true positives for 'fear' and the least number of true positives for'surprise'. It also predicts fear as sadness, and for 'anger', all four emotions except surprise are spread out evenly. Fear shows the darkest heatmap result, implicating the highest number of true positives, but at the

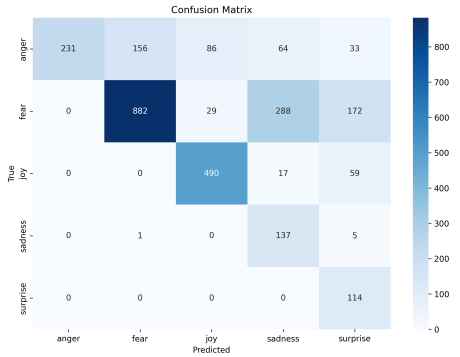

Figure 1: Confusion Matrix with k=5.

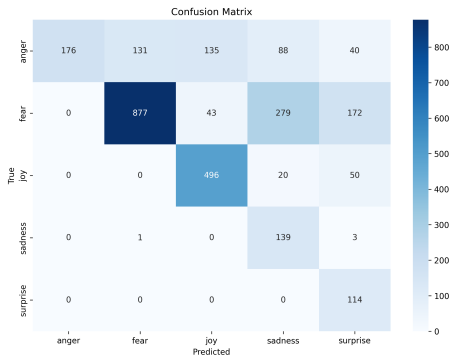

Figure 2: Confusion Matrix with k=10.

same time, the model predicts fear as sadness or surprise as a mistake. These results may be so because of the lack of training text to let the model about 'anger' or 'fear', but at the same time, it might be that the emotion 'anger' or 'fear' itself is not a single feature of emotion but a mixture of multiple emotions. Similarly, fear comes together with sadness or surprise [11]

5 SHAP : SHapley Additive Explanations

SHAP provides a unified approach to feature importance by calculating the contribution of each feature to the model's predictions. By visualizing SHAP values, we gain insights into how individual features impact the model's decision-making process, enhancing model transparency and interpretability. [9]

5.1 Visualize the impact on all the output classes

The Figure 3 visualizes the impact on a single class, here 'anger'. We slice out just a category 'anger' to visualize the model output towards that emotion, showing what keywords the model relies on the most to classify text into specific emotions. The first sentence likely represents anger, given the strong SHAP values and the high probability.The second sentence is associated with a weaker probability, meaning its words have less decisive influence.The third sentence has minimal impact, suggesting it might be more neutral or not strongly aligned with a specific category.

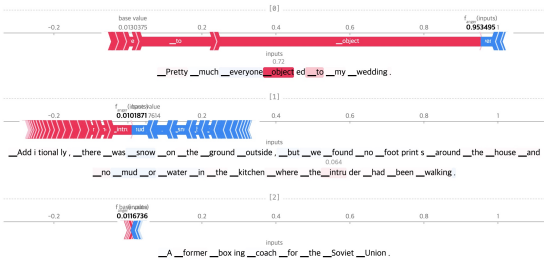

Figure 3: Shap-anger

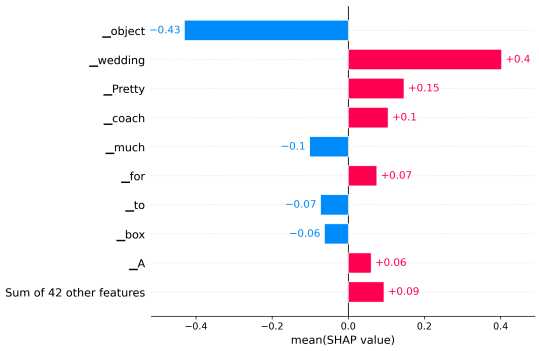

Figure 4: Shap-joy

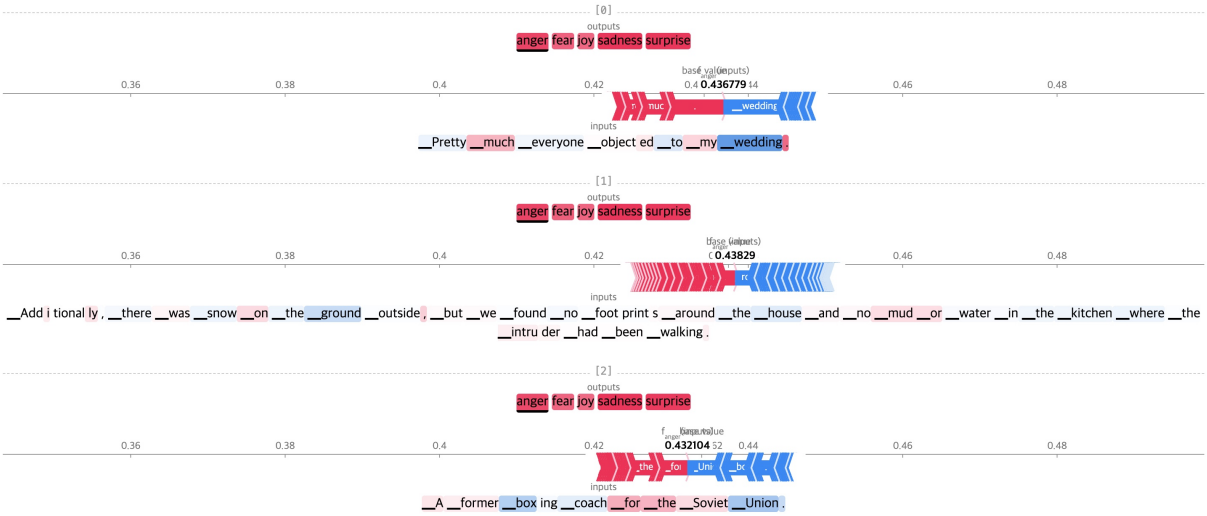

Figure 5: log-odds of Joy

5.2 Plotting the top words impacting a specific class

Figure 4 represents the SHAP values for individual words in their contribution to the "joy" class prediction across three examples. The model is balancing both positive and negative signals to determine whether a given text expresses joy. "*wedding*" is the most influential word in classifying joy, which makes sense as weddings are often associated with happiness. "*object*" strongly decreases joy, possibly because it introduces a more neutral or abstract context rather than an emotional one.

5.3 Explain the log odds instead of the probabilities

The Figure 5 represents a log-odds analysis of emotion classification on text, showing how specific words influence different emotions. This helps identify biases or unexpected classifications in sentiment analysis models. Each sentence has a base emotion probability (e.g., 0.43678), indicating its alignment with an emotional category. Higher log-odds values indicate stronger alignment with certain emotions. Words in red or pink increase the likelihood of negative emotions, while words in blue may reduce it or shift towards neutral or positive emotions. For instance, in 'Pretty much everyone objected to my wedding,' 'objected' contributes strongly to negative emotions like anger and sadness, while 'wedding' has a neutral or positive impact, making this part of the text stand out as odd compared to the rest of the sentence.

6 Discussion

RoBERTuna shows relatively fair performance on emotion detection in text with an F1 score of 0.951 and accuracy of 0.952 by leveraging Optuna for automated search of the optimal hyperparameters and the K-fold cross validation technique to derive the best utility of the limited amount of dataset. The model correctly predicts 'fear' the most, followed by 'joy'. But it also makes mistakes by categorizing an emotion to the others. To increase the accuracy, the model needs to get trained with larger datasets with colloquial or scholarly texts. Moreover, additional inputs such as conversation background, or speaker information may provide a positive impact to enhance the model. Further performance improvement is possible by asserting LLM-as-a-judge to the training process as a reward-penalty metric, where evaluation scores are adjusted based on the comparison between the model's emotion detection results and those of the chosen LLM.

7 Contribution

Tasks are done fairly between the two authors. Contribution statistics on Git repository does not reflect the individual work but the entire team.

References

[1] Md. Ali Akber, Tahira Ferdousi, Rasel Ahmed, Risha Asfara, Raqeebir Rab, and Umme Zakia. Personality and emotion—a comprehensive analysis using contextual text embeddings. *Natural Language Processing Journal*, 9:100105, 2024.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019.

[3] Sieun An, Li-Jun Ji, Michael Marks, and Zhiyong Zhang. Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. *Frontiers in psychology*, 8:610, 2017.

[4] Paweł Bujnowski, Bartłomiej Kuzma, Bartłomiej Paziewski, Jacek Rutkowski, Joanna Marhula, Zuzanna Bordzicka, and Piotr Andruszkiewicz. Samsemo: New dataset for multilingual and multimodal emotion recognition. In *Proc. Interspeech 2024*, pages 2925–2929, 2024.

[5] Gianluca Calò, Francesco Massafra, Berardina De Carolis, and Corrado Loglisci. Emotion hunters at emit: Categorical emotion detection combining bert and chatgpt models. 2023.

[6] Wikipedia contributors. Robert plutchik — Wikipedia, year = 2024, url = https://en.wikipedia.org/w/index.php? title=Robert$_P lutchikoldid = 1250459695, note =$ "[$online; accessed 31 - january - 2025$]".

[7] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009, 2021.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[9] Scott Lundberg and Su-In Lee. Emotion classification multiclass example, 2024. Accessed: Feb 2, 2025.

[10] Abdullah Al Maruf, Fahima Khanam, Md. Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*, 12:18416–18450, 2024.

[11] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297, 2013.

[12] Chunting Wan, Chuanpei Xu, Dongyi Chen, Daohong Wei, and Xiang Li. Emotion recognition based on a limited number of multimodal physiological signals channels. *Measurement*, 242:115940, 2025.