tinaal at SemEval-2025 Task 11: Enhancing Perceived Emotion Intensity Prediction with Boosting Fine-Tuned Transformers

Ting Zhu and **Liting Huang** and **Huizhi Liang**
Newcastle University, Newcastle Upon Tyne, England
T.Zhu11, L.Huang29, huizhi.liang@newcastle.ac.uk

Abstract

This paper presents a framework for perceived emotion intensity prediction, focusing on SemEval-2025 Task 11 Track B. The task involves predicting the intensity of five perceived emotions—anger, fear, joy, sadness, and surprise—on an ordinal scale from 0 (no emotion) to 3 (high emotion). Our approach builds upon our method introduced in the WASSA workshop and enhances it by integrating ModernBERT in place of the traditional BERT model within a boosting-based ensemble framework. To address the difficulty in capturing fine-grained emotional distinctions, we incorporate class-preserving mixup data augmentation, a custom Pearson CombinLoss function, and fine-tuned transformer models, including ModernBERT, RoBERTa, and DeBERTa. Compared to individual fine-tuned transformer models (BERT, RoBERTa, DeBERTa and ModernBERT) without augmentation or ensemble learning, our approach demonstrates significant improvements. The proposed system achieves an average Pearson correlation coefficient of 0.768 on the test set, outperforming the best individual baseline model. In particular, the model performs best for sadness (r = 0.808) and surprise (r = 0.770), highlighting its ability to capture subtle intensity variations in the text. Despite these improvements, challenges such as data imbalance, performance on low-resource emotions (e.g., anger and fear), and the need for refined data augmentation techniques remain open for future research.

1 Introduction

Emotions play a critical role in human communication, influencing decision-making, relationships, and interactions. In recent years, automatic detection and modelling of emotions in the text have attracted significant attention within the natural language processing (NLP) community due to their potential applications in areas such as mental health support, and personalized recommender systems (Zad et al., 2021). Despite this progress, emotion recognition remains challenging because of the complexity and subjectivity inherent in emotional expression. This study focuses on the detection of perceived emotions, which predicts the emotions that most people would associate with a given text. Perceived emotions are influenced by culture and individual differences, making their detection complex and nuanced (Van Woensel, 2019). Previous studies, such as those of Mohammad et al. (Mohammad et al., 2018), have highlighted the importance of perceived emotions in tasks like sentiment analysis and emotion intensity prediction.

In this work, we explore *SemEval-2025 Task 11 Track B: Emotion Intensity Prediction* of the shared task on Knowledge Representation and Reasoning (Muhammad et al., 2025b). Track B aims to predict the intensity of perceived emotions, joy, sadness, fear, anger, surprise, or disgust, on an ordinal scale ranging from 0 (no emotion) to 3 (high emotion). Previous research has explored techniques such as ordinal regression (Mehta et al., 2019; Yang et al., 2024) and multi-task learning (Akhtar et al., 2019) to better model emotion intensity. However, there remains a gap in optimizing these models specifically for emotion intensity prediction, particularly in learning from limited data, preserving ordinal relationships between intensity levels, and integrating augmentation techniques that improve generalization without distorting emotional meaning.

To address these challenges, this study introduces a novel framework that builds on our method proposed in the WASSA workshop, enhancing it with ModernBERT as a replacement for the traditional BERT model. The proposed approach integrates class-preserving mixup data augmentation, a custom Pearson CombinLoss function, and a boosting-based ensemble strategy to improve the model's ability to handle the ordinal nature of intensity labels, capture subtle emotional variations, and enhance robustness across different contexts.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 302–307

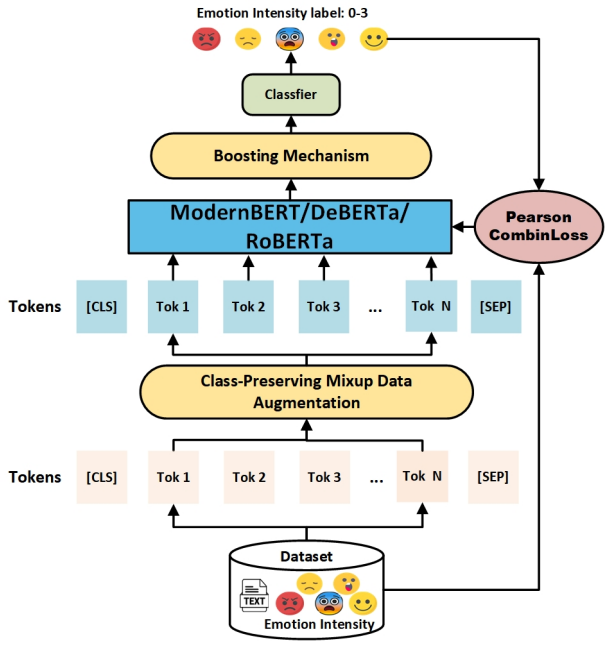

Figure 1: Overview of the Proposed Emotion Intensity Prediction Framework. The system integrates Class-Preserving Mixup Data Augmentation, Pearson CombinLoss, and a Boosting Ensemble with fine-tuned ModernBERT, RoBERTa, and DeBERTa models.

The ensemble leverages multiple fine-tuned transformer models, including ModernBERT, RoBERTa, and DeBERTa, to combine their strengths and improve generalization. Through several evaluations on SemEval-2025 Task 11 Track B, we demonstrate that our method significantly outperforms individual transformer-based models, achieving improved Pearson correlation scores, particularly for sadness and surprise.

2 Methodology

Our approach enhances emotion intensity prediction by integrating ModernBERT into a boosting-based ensemble framework, based on the method introduced in the WASSA workshop (Huang and Liang, 2024). Figure 1 illustrates our method for Track B in SemEval-2025 Task 11. It consists of data augmentation, the Pearson CombinLoss function, fine-tuned ModernBERT, DeBERTa or RoBERTa models, and the effective boosting strategy. In the following, we describe each component in detail.

2.1 Class-Preserving Mixup Data Augmentation

Traditional mixup methods (Smucny et al., 2022) is a widely used regularization technique that improves model generalization by blending samples across all classes. However, emotion intensity prediction introduces issues such as semantically unrealistic emotion blending, misalignment with the ordinal structure of intensity labels, and imbalance in augmented data distribution. Therefore, this paper proposes class-preserving mixup data augmentation. This method ensures that only samples within the same emotion category are mixed. This preserves the semantic integrity of the text while introducing controlled variation, allowing the model to better generalize across different expressions of the same emotion. Mathematically, given an input sequence X_i, and its label y_i, a mixed sequence $\tilde{X}_i$ is generated by selectively replacing a span of X_i with another sample X_k from the same class:

$$\tilde{X}_i[j] = \begin{cases} X_i[j], & \text{if } j \notin [s, e] \\ X_k[j], & \text{if } j \in [s, e] \end{cases} \quad (1)$$

where $[s, e]$ is the randomly selected span, and X_k is a randomly chosen sample from the same class as X_i. In emotion classification tasks, data augmentation enhances model performance by generating new training data. For example, mixing two anger-related texts— text 1 ("*One jumped on my leg and bit me.*") and text 2 ("*I finally managed to sit up on my hands and knees, and glaring up at her.*")—using Class-Preserving Mixup (with alpha = 0.1) produces: "*One jumped on my leg and bit me, and I finally managed to sit up, glaring up at her with anger.*" This mixed text retains the original narratives while strengthening emotional coherence. A corresponding image (see Figure 2) visualizes the process, showing an animal biting a person's leg and another person sitting up, glaring defiantly. Such augmentation enriches data diversity and improves emotion classification accuracy.

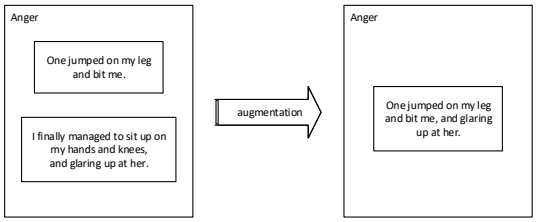

Figure 2: Example of Class-Preserving Mixup Augmentation in the Anger Category. Two original anger-related text snippets are combined by replacing a span in the first sentence with content from the second.

2.2 Pearson CombinLoss Function

To improve the performance of the model and take into account the layer-wise penalty, we use a new loss function that combines three parts: Cross-Entropy Loss, Structured Contrastive Loss, and Pearson Correlation Loss. We also introduce a hierarchical penalty matrix $\mathbf{P}$ to capture penalties for incorrect predictions based on the difference between predicted and true classes. The matrix is computed as:

$$P_{i,j} = \exp(-\gamma|i-j|) \tag{2}$$

where i and j are class indices, and γ controls the sharpness of penalty. This matrix penalizes predictions more severely, as they are further away from the true class.

This Pearson CombinLoss Function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{combined}} &= \alpha\mathcal{L}_{\text{CE}} + \gamma\mathcal{L}_{\text{SC}} + \beta\mathcal{L}_{\text{P}} \\ &= \alpha\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(y_i \mid \mathbf{x}_i)\right) \\ &\quad + \gamma\left(-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}P_{y_i,j}\log p(j \mid \mathbf{x}_i)\right) \\ &\quad + \beta\left(-\frac{\text{Cov}(\hat{\mathbf{y}},\mathbf{y})}{\sqrt{\text{Var}(\hat{\mathbf{y}})\cdot\text{Var}(\mathbf{y})}}\right)\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ is standard cross-entropy loss, $\mathcal{L}_{\text{SC}}$ is structured contrastive loss incorporating hierarchical penalties, and $\mathcal{L}_{\text{P}}$ is pearson correlation loss. α, γ, β are hyperparameters that control the weight of each component. $p(y_i \mid \mathbf{x}_i)$ is the predicted probability of the true class y_i for input $\mathbf{x}_i$. $P_{y_i,j}$ is the penalty for predicting class j when the true class is y_i, and C is the total number of classes. $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the predicted and true distributions, respectively, and Cov and Var denote covariance and variance.

2.3 Boosting Technique

Boosting (Tyralis and Papacharalampous, 2021) is employed to enhance the robustness and performance of the system by utilizing an ensemble strategy combined with weighted averaging. This approach capitalizes on the strengths of individual models to produce more accurate and reliable final predictions. The methodology involves using the Pearson Correlation Coefficients of the models as weights, thereby adjusting the final model output to optimize performance. This boosting mechanism ensures that the final predictions leverage the strengths of individual models, weighted by their respective Pearson Correlation Coefficients, resulting in improved performance across evaluation metrics.

$$\hat{y}_{\text{final}} = \frac{\sum_{i=1}^{M} w_i\hat{y}_i}{\sum_{i=1}^{M} w_i}, \quad w_i = r_i \tag{4}$$

where M is number of models in the ensemble and $\hat{y}_i$ represents predictions from the i-th model. r_i is Pearson Correlation Coefficient for the i-th model. w_i is weight assigned to the i-th model based on r_i.

3 Experiments and Results

3.1 Datasets

This study focuses on *SemEval-2025* Task 11 Track B: Emotion Intensity from the competition, which involves predicting the intensity of perceived emotions for text snippets. The task requires determining the degree of intensity for several perceived emotions: *joy*, *sadness*, *fear*, *anger*, *surprise*, or *disgust*. Emotion intensities are categorized into four ordinal classes: 0 (No emotion), 1 (Low emotion), 2 (Moderate emotion), and 3 (High emotion), reflecting the perceived degree of emotional expression in the text. These classes provide a structured scale to assess the intensity of emotion.

The dataset used in this study is a subset of the competition data (Muhammad et al., 2025a), focusing only on *English* language. The English dataset consists of training, development and test sets. Each sample contains: A unique identifier (*ID*), A short text snippet (*text*), Intensity scores for the five emotions (*anger, fear, joy, sadness, surprise*). Table 1 is an example of a training sample.

Table 1: Example of a Training Sample

Text: *"Then the screaming started."*

Anger	Fear	Joy	Sadness	Surprise
0	3	0	1	2

Other Information: ID: `eng_train_00001`.

The dataset consists of three subsets: a Training Set with 2,768 samples containing both textual data and labeled emotion intensities, a Development Set with 116 samples for fine-tuning and validation,

Table 2: Dataset Summary with Emotion Counts (Non-zero Counts) and Mean Scores

Dataset	Anger	Fear	Joy	Sadness	Surprise	Total
Train	333	1611	674	878	839	2768
Dev	16	63	31	35	31	116
Test	–	–	–	–	–	2767
Mean_Train	0.18	0.93	0.35	0.50	0.41	–
Mean_Dev	0.23	0.83	0.37	0.47	0.37	–

Note: Emotion sizes represent the non-zero counts of samples for each emotion category.

Table 3: Mixup-Augmented Dataset Summary with Non-zero Emotion Counts

Dataset	Anger	Fear	Joy	Sadness	Surprise	Total
Train-mixup	668	3224	1350	1758	1680	5536
Dev-mixup	34	128	64	72	64	234
Test-mixup	–	–	–	–	–	5536

and a Test Set with 2,767 samples containing only textual data for evaluating model predictions. The average emotion scores for the training set, presented in Table 2, reveal an imbalance. The development set follows the same structure as the training set, with similar mean emotion intensities, while the test set is unlabeled, emphasizing its role in assessing model performance. After applying mixup augmentation, the updated emotion distributions are detailed in Table 3.

3.2 Evaluation Metric

The official evaluation metric for Track B: Emotion Intensity task in this competition was the Pearson Correlation Coefficient, which measured the correlation between the gold-standard labels and the predicted ones. The Pearson Correlation Coefficient r ranges from -1 to 1, where values closer to 1 indicate a stronger positive correlation between the predicted and gold-standard values. The coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (5)$$

where $\hat{y}_i$ and y_i represent the predicted and gold-standard labels, respectively, $\bar{\hat{y}}$ and $\bar{y}$ are their means, and n is the number of samples.

3.3 Experiment Setup

The class-preserving mixup technique and Pearson CombineLoss Function were integrated into the training pipeline for state-of-the-art language models such as ModernBERT, RoBERTa, and DeBERTa. All models were implemented with the PyTorch framework, ensuring seamless integration with transformer architectures. The experiments were conducted on NVIDIA A4000 GPUs. The training process utilized the Adam optimizer with exponential decay (decay factor $\gamma = 0.99$), a batch size of 32. The models were trained and evaluated for each emotion category using different hyperparameter configurations. Table 4 presents the corresponding hyperparameters used for each emotion. Each model was trained on the emotion intensity prediction task using different configurations for the loss function parameters $\alpha, \beta, \theta = 0.8$. The learning rate (LR) and number of epochs were adjusted to optimize performance for each emotion category.

Table 4: Hyperparameters and Pearson Correlation Results for Each Emotion

Emotion	α	β	LR	Epochs
Anger	1	0.8	4×10^{-5}	6
Fear	0.5	0.5	4×10^{-5}	5
Joy	0.1	0.9	1×10^{-5}	10
Sadness	0.2	0.8	1×10^{-5}	8
Surprise	0	1.0	4×10^{-5}	9

Table 5: Pearson Correlation Coefficient (r) for each emotion on the development and test sets

Emotion	Dev Set Pearson	Test Set Pearson
Anger	0.656	0.760
Fear	0.780	0.735
Joy	0.820	0.768
Sadness	0.747	0.808
Surprise	0.738	0.770
Average	**0.748**	**0.768**

3.4 Results and Discussions

Our proposed framework integrates ModernBERT, Class-Preserving Mixup, Pearson CombinLoss, and a Boosting-Based Ensemble Strategy, achieving significant improvements in perceived emotion intensity prediction. By applying Class-Preserving Mixup and Pearson CombinLoss, we observed further performance gains. The augmentation strategy ensured label consistency, while the loss function helped the model better capture the ordinal nature of intensity levels. The boosting ensemble strategy further enhanced performance by leveraging the strengths of multiple models, leading to a fi-

Table 6: Pearson Correlation Coefficient (r) for Each Model on dev set

Model	Augmentation	Pearson	Anger	Fear	Joy	Sadness	Surprise	AVG.
BERT	×	×	0.521	0.708	0.767	0.677	0.605	0.656
DeBERTa	×	×	0.606	0.725	0.744	0.698	0.645	0.684
RoBERTa	×	×	0.546	0.715	0.772	0.631	0.690	0.670
ModernBERT	×	×	0.538	0.709	0.694	0.683	0.609	0.647
BERT	✓	×	0.569	0.702	0.772	0.661	0.609	0.663
DeBERTa	✓	×	0.656	0.732	0.803	0.722	0.691	0.717
RoBERTa	✓	×	0.564	0.746	0.788	0.700	0.690	0.700
ModernBERT	✓	×	0.605	0.713	0.724	0.701	0.653	0.679
BERT	✓	✓	0.615	0.703	0.770	0.687	0.646	0.684
DeBERTa	✓	✓	0.631	0.735	0.797	0.722	0.705	0.718
RoBERTa	✓	✓	0.600	0.740	0.783	0.708	0.702	0.707
ModernBERT	✓	✓	0.604	0.718	0.737	0.714	0.665	0.688
Boosting(Ours)	✓	✓	**0.656**	**0.780**	**0.820**	**0.747**	**0.738**	**0.748**

nal average Pearson correlation of 0.768 [1] on the test set. (Table 5). The ensemble demonstrated the strongest performance on sadness (r = 0.808) and surprise (r = 0.770), indicating its ability to effectively capture subtle intensity variations.

The performance of individual transformer models, as well as the ensemble results, is summarized in Table 6 on development set. Among the individual models, DeBERTa achieved the highest overall Pearson correlation coefficient ($r = 0.718$), with strong performance for *joy* ($r = 0.803$) and *fear* ($r = 0.735$). RoBERTa followed with an overall Pearson correlation of 0.707, performing best for *joy* ($r = 0.788$) and *fear* ($r = 0.740$). BERT showed the lowest overall performance ($r = 0.684$), particularly struggling with *anger* ($r = 0.615$). These results highlight the varying capabilities of individual transformer models in capturing the nuances of emotional intensities.

Despite these improvements, challenges remain to accurately model fear and anger, where the system's performance was relatively lower. This is likely due to data imbalance, where fewer training samples for these emotions limited the model's ability to generalize. Furthermore, fear and anger often depend on nuanced contextual cues, which may not be fully captured by current transformer-based models.

[1]This score is based on a post-evaluation run using the test set after the gold labels were released. The leaderboard score (0.67) corresponds to an earlier submission.

4 Conclusions

This paper presents a novel framework for perceived emotion intensity prediction, developed for SemEval-2025 Task 11 Track B, by integrating ModernBERT within a boosting-based ensemble model. The proposed approach builds upon the method introduced in the WASSA workshop and incorporates Class-Preserving Mixup data augmentation, a custom Pearson CombinLoss function, and fine-tuned transformer models to address key challenges such as the ordinal nature of emotion intensity labels, and capturing fine-grained emotional distinctions. Our system achieved an average Pearson correlation coefficient of 0.768 on the test set, demonstrating significant improvements over baseline models. The ensemble approach, leveraging ModernBERT, RoBERTa, and DeBERTa, was particularly effective in modeling subtle variations in sadness and surprise, achieving Pearson correlations of 0.808 and 0.770, respectively. However, challenges remain in accurately modeling *fear* and *anger*, likely due to limited training samples and inherent subjectivity in emotional expression.

Future work could explore data augmentation strategies, adaptive loss functions tailored to ordinal emotion scales, and context-aware transformer models to enhance emotion intensity prediction further. Additionally, integrating multimodal signals such as prosody and speech patterns could help improve robustness in real-world applications, particularly in mental health support and personalized conversational AI.

References

Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*.

Liting Huang and Huizhi Liang. 2024. Zhenmei at wassa-2024 empathy and personality shared track 2 incorporating pearson correlation coefficient as a regularization term for enhanced empathy and emotion prediction in conversational turns. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 399–403.

Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Y Javaid. 2019. Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors*, 19(8):1897.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, et al. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Jason Smucny, Ge Shi, Tyler A Lesh, Cameron S Carter, and Ian Davidson. 2022. Data augmentation with mixup: Enhancing performance of a functional neuroimaging-based prognostic deep learning classifier in recent onset psychosis. *NeuroImage: Clinical*, 36:103214.

Hristos Tyralis and Georgia Papacharalampous. 2021. Boosting algorithms in energy research: A systematic review. *Neural Computing and Applications*, 33(21):14101–14117.

Lieve Van Woensel. 2019. What if your emotions were tracked to spy on you?

Huiyu Yang, Liting Huang, Tian Li, Nicolay Rusnachenko, and Huizhi Liang. 2024. hyy33 at wassa 2024 empathy and personality shared task: Using the combinedloss and fgm for enhancing bert-based models in emotion and empathy prediction from conversation turns. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 430–434.

Samira Zad, Maryam Heidari, H James Jr, and Ozlem Uzuner. 2021. Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0255–0261. IEEE.