Dianchi at SemEval-2025 Task 11: Multilabel Emotion Recognition via Orthogonal Knowledge Distillation

Zhenlan Wang, Jiaxuan Liu and **Xiaobing Zhou***
School of Information Science and Engineering,
Yunnan University, Kunming 650091, China
*Corresponding author: zhouxb@ynu.edu.cn

Abstract

This paper presents our team's approach in SemEval-2025 Task 11: "Task 11: Bridging the Gap in Text-Based Emotion Detection", which aims to predict the speaker's perceived emotions in given target text segments(Muhammad et al., 2025). Our methodology employs a BERT pre-trained model for text processing combined with knowledge distillation and a dynamic data expansion approach. After initializing training parameters, we train student models by calculating classification and distillation losses for parameter updates. New prediction data is generated through periodic evaluation and incorporated into the original dataset to update the data loader for enhanced data augmentation, while synchronously updating both teacher and student model weights. Our system achieved an accuracy of 0.7 in the English multi-label text classification task in Subtask A of SemEval-2025 Task 11. The code is available at https://github.com/w2060772766/a1.

1 Introduction

Multi-label text classification advances beyond the semantic unidimensionality limitation of traditional single-label classification by assigning multiple interrelated labels to a single text, thereby effectively capturing the complexity of coexisting emotions in real-world scenarios (Zheng et al., 2024). As a pivotal technology for revealing human cognitive states, Emotion Recognition (ER) demonstrates significant application value across diverse domains (Alaluf and Illouz, 2019; Muhammad et al., 2025), including consumer behavior analysis (Abdul-Mageed et al., 2018) and community mental health monitoring (Volkova and Bachrach, 2016). However, existing methods often suffer from misdetection of fine-grained emotional features due to insufficient long-range semantic modeling capabilities, while also facing overfitting risks in small-scale annotated data scenarios.

To address these challenges, this study proposes an innovative knowledge distillation framework: (1) By leveraging hidden-layer representations of pre-trained language models as soft target supervision signals, we enhance the capture of deep semantic correlations through a teacher-student parameter transfer mechanism; (2) A pseudo-label extension strategy is integrated with dynamic data augmentation to mitigate distributional shift issues during training. This approach inherits large-scale models' powerful semantic encoding capabilities while enabling robust multi-label emotion inference through a lightweight architecture, thereby providing novel insights for fine-grained emotion detection in complex real-world environments.

2 Background

In the field of Natural Language Processing, BERT—a Transformer-based pre-trained language model (Devlin et al., 2019)—has significantly enhanced textual representation capabilities through its pre-training paradigm of masked language modeling and next-sentence prediction. However, the substantial parameter size of BERT models incurs high computational costs, limiting their industrial deployment. To address this, Knowledge Distillation (KD) has been introduced for model lightweighting. Original KD (Hinton et al., 2015) operates by transferring implicit knowledge—such as output-layer probability distributions and intermediate-layer attention weights—from complex teacher models to compact student models.

In the context of BERT compression, Sanh (Sanh et al., 2019) developed DistilBERT, which

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1108–1112
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

reduces the model size by 40% while retaining 97% of the original performance through layer reduction and a teacher-student attention alignment loss function. Subsequent work has extended this framework: Jiao et al. (Jiao et al., 2019) proposed TinyBERT, which incorporates attention matrix mapping and hidden state adaptation to enable layer-wise knowledge transfer, achieving an accuracy gap of merely 3% compared to BERT-base on GLUE benchmarks. Notably, knowledge distillation applications in BERT optimization have evolved beyond single-model compression to innovative directions such as multimodal pre-training (Sun et al., 2019) and dynamic architecture pruning, demonstrating its enduring potential to balance model efficiency and performance.

3 System Overview

In this section, we delineate our methodological framework. Our approach leverages the BERT pre-trained model for text sequence processing and contextual representation learning. We synergistically integrate Knowledge Distillation (KD) with advanced textual data augmentation strategies to enhance generalization performance.

3.1 Pre-training

We employ the BERT pre-trained model for text classification. The input text is first tokenized into subword units using BERT's tokenizer and mapped to numerical IDs through its pre-trained vocabulary. The tokenized sequence undergoes hierarchical feature extraction via BERT's multi-layer self-attention mechanisms and feedforward neural networks, generating high-dimensional contextual embeddings(Vaswani et al., 2017). To structure the training data, three containers are initialized: ids for sample identifiers, texts for original text content, and labels for emotion category annotations. During batch iteration, each sample's identifier, text, and label are sequentially appended to their respective containers. For classification, a task-specific fully connected layer is appended to the BERT architecture. During inference, input_ids $tokenIDs$ and attention_mask (sequence padding indicators) are fed into the model to extract final-layer representations. Crucially, the feature vector corresponding to the [CLS] token is leveraged as the aggregated semantic signal to predict emotion categories through the classifier, enabling robust textual emotion analysis within an end-to-end framework.

3.2 Knowledge Distillation (KD)

Knowledge Distillation (KD) is a technique that transfers knowledge from a teacher model to a student model(Ma et al., 2024), aiming to enhance the student's performance or reduce its computational footprint while maintaining high accuracy. In our implementation, both the teacher and student models share an identical BERT architectural structure but are initialized with independent parameters. The teacher model is trained directly on the original task, while the student model is optimized to mimic the teacher's knowledge while retaining lightweight computational demands.

During training, a composite loss function is designed to guide the student model. We integrate a classification loss (task-specific supervision) with a distillation loss (knowledge transfer regularization), mediated by a balancing parameter λ to harmonize their contributions:

$$\mathcal{L}_{\text{total}} \ = L_{\text{CE}} \ + \lambda L_{\text{KL}} \tag{1}$$

To ensure the student model can make accurate predictions, we adopt a classification loss function, using cross-entropy loss to measure the discrepancy between predicted probabilities and ground-truth labels. The conventional binary cross-entropy loss formula is expressed as:

$$\mathcal{L}_{\text{CE}_\text{o}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{2}$$

Normalize the weight vectors so that each w_i satisfies $\|w\| = 1$. Compute the inner product matrix of the normalized weights:

$$\mathcal{S}_{ij} = W_i^* W_j^* \tag{3}$$

Take the upper triangular part (excluding the diagonal) and sum the positive inner product values:

$$\mathcal{L}_{\text{sim}} = \sum_{i<j} \mathcal{S}_{ij} \cdot I(\mathcal{S}_{ij} > 0) \tag{4}$$

Finally, the total loss is obtained as:

$$\mathcal{L}_{\text{CE}} = \mathcal{L}_{\text{CEo}} + \lambda \cdot \mathcal{L}_{\text{sim}} \tag{5}$$

To enable the student model to learn the "soft" knowledge from the teacher model, i.e., the semantic information embedded in its probability

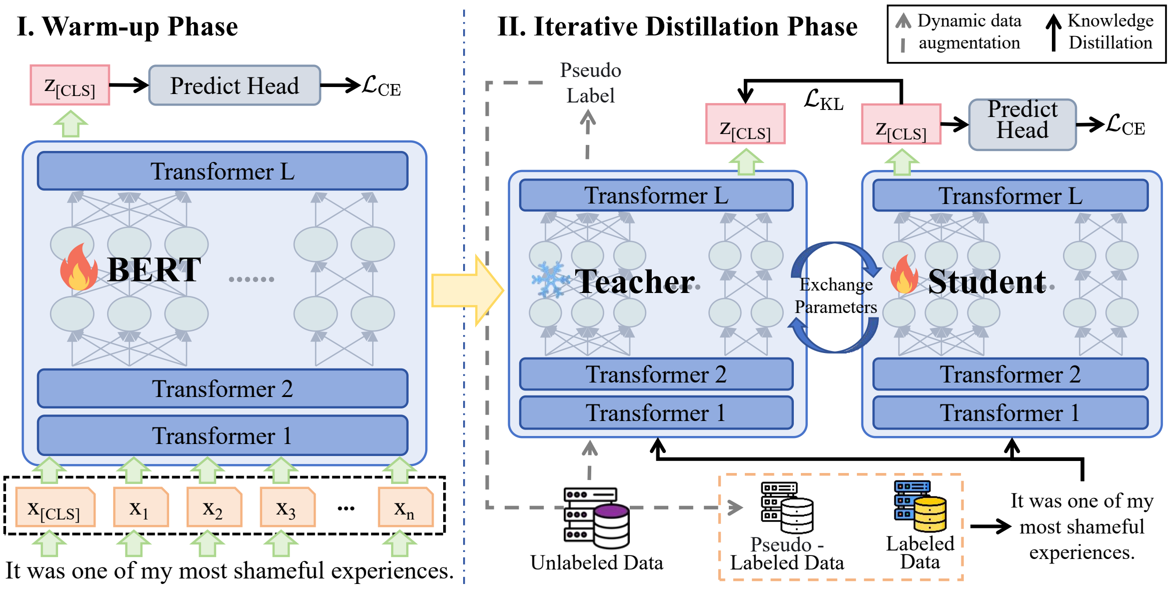

Figure 1: System architecture: left side of the figure is our pre-trained model, and the right side is the distillation model we employed.

distributions, we employ the Kullback-Leibler divergence (KL divergence) to quantify the discrepancy between the probability distributions of the student and teacher models(Li et al., 2023). A temperature parameter T is introduced to smooth out the sharp probability distributions, thereby enabling the student model to better learn the soft knowledge and semantic representations from the teacher model, ultimately enhancing the model's performance and generalization capability:

$$\mathcal{L}_{\text{KL}} = \frac{1}{T^2} KL(p^{\text{t}} \| p^{\text{s}}) \tag{6}$$

The KL divergence is defined as:

$$KL(p^{\text{t}} \| p^{\text{s}}) = \sum_{c=1}^{C} p_c^{\text{t}} \cdot \log\left(\frac{p_c^{\text{t}}}{p_c^{\text{s}}}\right) \tag{7}$$

where p_c^t represents the predicted probability value of the teacher model for the c class, and p_c^s denotes the predicted probability value of the student model for the c class.

The knowledge distillation technique in my work demonstrates three key advantages over conventional methods: Firstly, the soft labels generated by the teacher model (enhanced through temperature scaling) effectively transfer implicit correlations between multi-label emotions, overcoming the semantic rigidity of traditional hard labels (binary 0/1 supervision) to capture compound emotional features. Secondly, the dynamic pseudo-label augmentation mechanism periodically integrates high-confidence prediction samples, expanding training data distribution while preserving multi-label semantic consistency, thereby avoiding the disruption of label co-occurrence relationships caused by traditional static augmentation methods. Finally, the alternating parameter update strategy between teacher and student models establishes an ”exploration-consolidation” cycle that preserves historical optimal knowledge while encouraging continuous optimization under new data distributions. Coupled with orthogonal constraints on classifier weights, this approach jointly resolves the feature space collapse issue induced by label co-occurrence in traditional methods, ultimately achieving enhanced precision and generalization in fine-grained emotion detection.

3.3 Model Training

During the training process, we first train the student model (s_model). Each epoch in the main training loop consists of training and evaluation phases. In the training phase, the student model performs a forward pass, calculates the combined loss (including classification loss and distillation loss), and updates model parameters through backpropagation. After specific epochs, s_model switches to evaluation mode to generate predictions on the validation set, which are then merged into the original training set to enhance data diversity. Each time the training set is expanded, the teacher model (t_model) inherits the weights from s_model, while s_model's weights are overwritten by those of a new_model (a fresh model initialized for continued learning on the updated training data). A step-wise scheduler is adopted for the learning rate, reduced every 10 epochs. Finally, predictions on the validation set are generated using the latest t_model and retained as results. The training process is shown in Figure 1.

4 Experimental Setup

This section details the configuration of Subtask A, including dataset structure and training strategies. The task data originates from the official CSV-formatted dataset released for SemEval 2024, comprising three splits: a training set, a development (dev) set, and a test set, each containing 2,768 annotated samples. Each sample follows a "text + sentiment label" structure: the text field contains the input sentence for classification, while the sentiment labels cover five fine-grained categories (anger, fear, joy, sadness, surprise) under a multi-label annotation scheme.

Through parameter tuning experiments, we identified optimal performance when training the model for 90 total epochs with a 15-epoch learning rate warm-up phase. The batch sizes were set to 32 for the training set and 128 for the validation set. During the warm-up phase (first 15 epochs), the initial learning rate was 3e-5, which decayed by 10% every 10 epochs. Additionally, a learning rate scheduler was implemented to facilitate model convergence and enhance performance.

5 Results and Analysis

5.1 Results

This section presents the results of our model for the English multi-label text classification task in Subtask A of SemEval-2025 Task 11. We compare our outcomes with the official benchmark data, using accuracy as the primary evaluation metric. Three experiments were conducted: (1) A baseline approach utilizing only BERT without knowledge distillation achieved an accuracy of 0.35; (2) When introducing distillation with data augmentation during preprocessing (replicating texts from underrepresented categories), performance significantly deteriorated, as shown in Table 1, where accuracy dropped from 0.7 to 0.68. Further analysis suggests that improper class balancing during augmentation may have disrupted the model's ability to generalize effectively.

5.2 Analysis

Firstly, compared to using the BERT method alone, knowledge distillation mitigates BERT's overfitting to dominant labels by softening the probability distribution of the teacher model(Oliver et al., 2018). However, in the third experiment, while naively replicating minority-class texts increased the dataset size, the mechanical duplication disrupted the complex co-occurrence relationships in multi-label samples (e.g., forcing an increased frequency of a specific label concurrently distorted the semantic distribution of other correlated labels). This introduced a cognitive bias in the model's perception of the true data distribution, thereby compromising the regularization benefits of distillation and impairing generalization capability.

Additionally, preprocessing steps may have inadvertently removed critical features or introduced distribution bias. The preprocessed data might also mismatch the input distribution of pre-trained models (e.g., BERT), compromising their semantic encoding capability. These factors could collectively degrade the final accuracy.

6 Conclusions

This paper details our participation in SemEval 2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, specifically subtask A (English). Our approach employs a BERT-based pretrained model to encode textual CLS token representations as features, which are then passed through a linear layer to generate logits for multi-label classification via sigmoid thresholding. We innovatively enhanced the cross-entropy (CE) loss by introducing orthogonal regularization on the fully connected layer's weight matrix (using an inverse distance penalty between weight vectors) and incorporated knowledge distillation during later training stages with a KL divergence constraint between the teacher and student model outputs. Throughout the training, validation set predictions were dynamically integrated into the training data every 15 epochs, alongside alternating parameter updates between the teacher and student models for iterative refinement. This

English							
approach	**anger**	**disgust**	**joy**	**sadness**	**surprise**	**macro f1**	**micro f1**
Without distillation	0.1026	0.5672	0.3182	0.4571	0..3284	0.3547	0.4181
With distillation	0.567	0.8085	0.7116	0.7057	0.6893	0.6964	0.7329
With data-preprocessing	0.5714	0.8	0.6154	0.7302	0.7	0.6834	0.7207

Table 1: The accuracy rates obtained without knowledge distillation, with knowledge distillation, and with preprocessing before distillation.

framework ultimately produced prediction files annotated with five emotion categories, combining regularization, dynamic augmentation, and distillation to address the challenges of multi-label emotion detection.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Yaara Benger Alaluf and Eva Illouz. 2019. Emotions in consumer studies. In *The Oxford Handbook of Consumption*, page 239. Oxford Univ. Press New York, NY, USA.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1504–1512.

Ying Ma, Xiaoyan Zou, Qizheng Pan, Ming Yan, and Guoqi Li. 2024. Target-embedding autoencoder with knowledge distillation for multi-label classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

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. 2025. 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.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578.

Guangmin Zheng, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2024. Instruction tuning with retrieval-based examples ranking for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*.