Neural Nexus at PsyDefDetect: Fine-Tuning RoBERTa with Focal Loss and Role-Tagged Dialogue History for Defense Level Detection

Subhrajyoti Basu
Heritage Institute of Technology
subhrajyoti479@gmail.com

Abstract

We describe our system for the PsyDefDetect shared task at BioNLP 2026, which focuses on classifying help-seeker utterances in multi-turn supportive conversations into nine psychological defense mechanism levels defined by the Defense Mechanism Rating Scales (DMRS). Our approach fine-tunes `roberta-base` using a composite training objective that combines focal loss, label smoothing, and square-root dampened class weights to address the severe label imbalance present in the PSYDEFCONV corpus, where the dominant class constitutes 52% of the training data. The input representation is constructed by concatenating up to eight dialogue turns with role-specific tags, separated using RoBERTa's native `</s>` tokens, followed by the target utterance marked using a `[TARGET]` token. Model selection is performed using macro-F1 based early stopping on a stratified 15% validation split, along with cosine learning rate decay for stable optimization. Our best submission achieves an official Leaderboard 1 (positive classes) macro-F1 score of **0.2556**, ranking 11th among 21 registered teams.

1 Introduction

Psychological defense mechanisms are unconscious strategies through which individuals manage anxiety and regulate emotional distress (Na et al., 2026b). In supportive conversations, a help-seeker's choice of defense—whether mature coping, intellectualization, or denial—shapes the trajectory of the interaction and the kind of support that is likely to be effective. Automatically detecting these defense levels is therefore relevant for clinical decision support, psychotherapy research, and emotionally intelligent dialogue systems. Despite this relevance, the problem remains largely unexplored in NLP.

The PsyDefDetect shared task at BioNLP 2026 (Na et al., 2026a) introduces this problem through the PSYDEFCONV dataset, the first conversational corpus annotated with DMRS-based defense levels. The task requires classifying help-seeker utterances into nine hierarchical categories using their dialogue context. This setting connects to ongoing work in psychotherapy-focused language modeling and mental health NLP (Na et al., 2025).

The task presents three main challenges. **(1) Severe class imbalance**: Level 7 (High-Adaptive Defenses) accounts for 52% of the training data, whereas Level 8 (Needs More Information) contains only 28 instances. **(2) Fine-grained clinical distinctions**: Adjacent DMRS levels often show strong surface overlap. For example, Disavowal (Level 3) and Obsessional defenses (Level 6) both involve verbal rationalization, but differ in whether the speaker avoids or intellectualizes affect. **(3) Context dependency**: The same utterance may correspond to different defense levels depending on the preceding dialogue, including both supporter and help-seeker turns.

To address these challenges, we adopt a focal loss objective with square-root dampened class weights, label smoothing, and a context-rich input representation that encodes up to eight dialogue turns with explicit role tags and RoBERTa separator tokens. Our best system achieves a private test macro-F1 score of 0.2556. We further provide per-class analysis and discuss the observed failure modes to inform future work.

2 Related Work

Liu et al. (2021) introduced the ESConv dataset and emotional support strategy classification, which also serves as the source corpus for PSYDEFCONV, while Na et al. (2025) highlight the gap between current NLP capabilities and real clinical requirements in psychotherapy-related applications. In imbalanced text classification, focal loss (Lin et al., 2017) has been widely used to focus

Proceedings of the BioNLP 2026 (Shared Tasks), pages 66–70

learning on harder and underrepresented examples, and Müller et al. (2019) show that it can work well with label smoothing by reducing overconfidence while preserving learning signal on difficult cases. More broadly, encoder-based dialogue models have shown strong performance in emotion recognition in conversation (Poria et al., 2019), where context dependency and class imbalance are also important challenges. Our input formulation follows this line of work by using role-tagged dialogue history with separator tokens to better capture conversational structure and dependencies.

3 Task and Data

The shared task is based on the PSYDEFCONV dataset (Na et al., 2026b), which is constructed from a stratified subset of the ESConv corpus of emotional support conversations (Liu et al., 2021). The dataset consists of 200 multi-turn dialogues between help-seekers and supporters, with a total of 2,336 help-seeker utterances annotated using the DMRS clinical framework by expert annotators.

Label schema. The task requires classifying each utterance into one of nine categories spanning the full DMRS hierarchy. This includes seven hierarchical defense levels along with two auxiliary categories: No Defenses", which corresponds to functional utterances without engagement in internal conflict, and Needs More Information", which captures ambiguous cases where the context is insufficient for reliable classification. Table 1 presents the full label distribution and corresponding training frequencies.

Class imbalance. The dataset exhibits significant class imbalance, which directly influences the modeling approach. Level 7 (High-Adaptive Defenses) is the dominant class constituting 52% of the data. In contrast, Levels 1, 2, 3, 4, 5 and 8 appear far less frequently which can be seen in Table 1. This level of imbalance makes standard fine-tuning difficult.

Evaluation. The shared task reports two official leaderboards based on macro-averaged F1 scores. Leaderboard 1 evaluates performance over the positive defense classes (Levels 1–8), while Leaderboard 2 evaluates performance across all classes (Levels 0–8). In both settings, macro-F1 is computed uniformly across classes, making performance on minority classes particularly important for overall system ranking.

Label	Defense Level	Count	%
0	No Defenses	296	15.9
1	Action	108	5.8
2	Major Image-Distorting	61	3.3
3	Disavowal	99	5.3
4	Minor Image-Distorting	84	4.5
5	Neurotic	48	2.6
6	Obsessional	172	9.2
7	High-Adaptive	968	52.0
8	Needs More Info	28	1.5
	Total	**1,864**	100

Table 1: Label distribution in the training set.

4 System Description

4.1 Input Representation

Each instance consists of a dialogue history and a target help-seeker utterance to classify. We concatenate the last k=8 dialogue turns, prefixing each turn with an explicit speaker role tag (`Seeker:` or `Supporter:`), and join turns with RoBERTa's native sentence-separator token `</s>` as a soft semantic boundary. The target utterance is appended after a final `</s>` under a `[TARGET] Seeker:` prefix that explicitly marks the utterance to classify:

```
Seeker: t1 </s> Supporter: t2
</s> ··· </s>
[TARGET] Seeker: target
utterance
```

All sequences are truncated and padded to a maximum of 256 tokens. Using k=8 turns—rather than the shorter windows (k=5) trialled in preliminary experiments—is important because defensive functioning often manifests as a *pattern* across multiple exchange rounds. For instance, escalating help-rejecting complaining (Level 1) is more detectable when the model observes a sequence of prior supporter suggestions being consistently deflected. The explicit `[TARGET]` marker prevents the model from attending ambiguously to any seeker turn in the history when classifying.

4.2 Pre-trained Model

We fine-tune `roberta-base` (Liu et al., 2019), a 12-layer Transformer encoder with 125M parameters, appending a two-layer classification head (dense + dropout + output projection) over the `<s>` (CLS) representation for 9-way classification. Hidden dropout and attention dropout are both set to 0.1 to regularize the encoder during fine-tuning on the small corpus.

In preliminary experiments, we also trained `distilbert-base-uncased` (Sanh et al., 2019) as a pilot baseline under a simpler setup with a shorter dialogue history window (k=5), standard inverse-frequency class weights, focal loss with γ=2.0, and no early stopping. The final `roberta-base` system improved over this pilot by approximately +0.04 macro-F1 on the official LB1 leaderboard. However, this improvement reflects not only the stronger encoder backbone but also the revised training and input configuration. RoBERTa's more aggressive pretraining strategy, including dynamic masking, larger batch training, and the removal of the next-sentence-prediction objective, likely contributes to richer contextual representations for the subtle pragmatic distinctions required by DMRS-level detection.

4.3 Training Objective

Class weights. To counteract majority-class dominance, we compute per-class inverse-frequency weights via scikit-learn's compute_class_weight ("balanced"):

$$w_c = \frac{N}{K \cdot n_c} \tag{1}$$

where N is total training samples, K=9, and n_c is the count of class c. Raw balanced weights are extreme for very rare classes (Level 8 yields $w \approx 7.4$), which destabilizes training by amplifying gradient noise from the few available samples. We apply square-root dampening to moderate the range while preserving the relative ordering:

$$\tilde{w}_c = \sqrt{w_c} \tag{2}$$

This keeps weights in the range ≈0.6–2.7, which we found empirically more stable than either undampened or cube-root dampening.

Focal loss with label smoothing. We combine focal loss (Lin et al., 2017) with label smoothing (Müller et al., 2019) to address both class imbalance and overconfidence simultaneously.

Label smoothing with ε=0.05 produces a soft target distribution:

$$\tilde{y}_c = \begin{cases} 1-\varepsilon & \text{if } c = y \\ \dfrac{\varepsilon}{K-1} & \text{otherwise} \end{cases} \tag{3}$$

The label-smoothed cross-entropy loss per sample i is:

$$\ell_i = -\sum_{c=1}^{K} \tilde{y}_c \log p_c^{(i)} \tag{4}$$

The focal weight down-weights confidently classified (high p_t) examples:

$$p_t^{(i)} = p_{y_i}^{(i)}, \qquad f_i = \left(1 - p_t^{(i)}\right)^{\gamma} \tag{5}$$

with γ=1.5. We use a lower γ than the original γ=2 (Lin et al., 2017) to avoid over-suppressing gradient signal from the dominant Level 7 class, which anchors shared encoder representations.

The final loss for a batch of B samples is:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \tilde{w}_{y_i} \cdot f_i \cdot \ell_i \tag{6}$$

This objective directs the model's representational capacity toward hard, often minority-class examples while preserving a meaningful learning signal from easier majority-class instances.

4.4 Training Configuration

We implement the custom objective in a subclassed HuggingFace `Trainer` (Wolf et al., 2020). The full hyperparameter configuration is given in Table 4 (Appendix A). Key settings:

- AdamW optimizer, learning rate 2×10^{-5}
- Cosine decay schedule, 6% linear warmup
- Weight decay 0.01, gradient clipping 1.0
- Batch size 16 / 32 (train / eval), fp16 precision
- Up to 12 epochs, early stopping with patience 3 on val macro-F1
- Stratified 85/15 train/validation split

At test time, logits are divided by temperature T=1.3 before the argmax to reduce overconfidence under the domain shift between training and test data (Guo et al., 2017).

5 Results

5.1 Validation Performance

Early stopping selected the best checkpoint at epoch 6 (of a maximum of 12), with validation macro-F1 of 0.288. Table 2 shows the per-class breakdown on the validation split (280 utterances). The model performs strongly on Level 0 (F1 = 0.832) and Level 7 (F1 = 0.677), which together constitute the majority of validation labels. Performance degrades sharply for low-frequency classes, with Levels 1, 3, and 5 showing very low or near-zero F1 scores, highlighting the difficulty of learning from extremely limited data.

L	Defense	P	R	F1
0	No Defenses	0.841	0.822	0.832
1	Action	0.059	0.063	0.061
2	Maj. Img.-Dist.	0.333	0.333	0.333
3	Disavowal	0.000	0.000	0.000
4	Min. Img.-Dist.	0.200	0.154	0.174
5	Neurotic	0.000	0.000	0.000
6	Obsessional	0.263	0.385	0.313
7	High-Adaptive	0.754	0.614	0.677
8	Needs More Info	0.167	0.250	0.200
	Macro avg.	0.291	0.291	0.288

Table 2: Per-class results on the validation split (280 utterances, 15% stratified split).

5.2 Private Test Performance

Table 3 compares our final submitted system against a preliminary DistilBERT pilot baseline. The final system achieves an official LB1 macro-F1 of 0.2556. Compared with the DistilBERT pilot run, this corresponds to an improvement of approximately +0.041 macro-F1. However, this comparison should be interpreted as a full-system comparison, since the pilot baseline used a different context window, loss configuration, and optimization setup as mentioned before.

System	Val F1	Official LB1 F1
DistilBERT pilot baseline	0.327	0.215
RoBERTa + focal (ours)	**0.288**	**0.256**

Table 3: Macro-F1 results on the validation split and the official LB1 test set. The DistilBERT pilot used a different training configuration, including a shorter context window (k=5), different loss hyperparameters, and no early stopping.

6 Analysis

Validation–private gap. Our system achieves a validation macro-F1 of 0.288 but only 0.256 on the private test set, leaving a gap of 0.032. We think this comes mainly from two factors. First, the validation split is small, with only 280 utterances, and some classes have fewer than ten samples, so even a few mistakes can move macro-F1 a lot. Second, repeated hyperparameter tuning across runs likely caused some implicit overfitting to this split, even though each run used fresh initializations. A more reliable setup would be stratified cross-validation or a separate validation set.

Failure modes on rare classes. The model struggles most on Levels 1, 3, and 5, which together have only 255 training samples. Data scarcity is one reason, but the bigger issue is that these classes often overlap with each other and with more frequent classes. In the validation set, Levels 3 and 5 collapse to F1 = 0.0, which shows how hard these labels are to learn from limited data. For example, a statement like *"nobody ever really understands me"* may fit Projection (Level 3, Disavowal) or Passive Aggression (Level 1, Action), depending on the surrounding context. These cases require a deeper understanding of the DMRS framework, not just surface cues. Level 8 is even harder, since it is defined by missing context rather than any clear positive signal, and 28 examples are simply not enough.

Effect of dialogue history window. Using a history window of k=8 turns worked better, especially for Action Defenses (Level 1) and Disavowal (Level 3). These defenses often appear as patterns across multiple turns. For example, repeated deflection of suggestions can signal Level 1, while more consistent engagement can align with Level 7. Shorter windows like k=5 miss these interaction patterns.

Temperature scaling. Temperature scaling with T=1.3 gave a small improvement on the private leaderboard. This is not surprising, since fine-tuned models on small domain-specific datasets often become overconfident. Temperature scaling is a simple fix, but it helps calibration (Guo et al., 2017).

7 Conclusion

We greatly appreciate the dataset as the strong class imbalance makes this task quite challenging. Although we attempted to address this issue, the performance still remains limited. In future work, we plan to explore more reliable ways of handling minority classes including controlled data augmentation methods such as paraphrasing and to incorporate DMRS-specific modeling techniques.

Limitations

Our system has several notable limitations. First, the small training set (1,864 utterances) and severe class imbalance fundamentally constrain fine-tuning approaches; Level 8 with 28 instances is practically unlearnable. Second, our models are encoder-only and operate without any explicit representation of the DMRS clinical framework; they learn surface correlates of defense levels rather than the underlying clinical constructs. Third, the

val/private gap (0.032) indicates unresolved overfitting to the specific validation split used throughout development. Finally, all experiments were conducted on a single GPU (NVIDIA T4, Google Colab), which precluded exploration of larger models or extensive ensemble methods due to time constraints.

Acknowledgments

We thank the PsyDefDetect organizing committee for preparing the PSYDEFCONV dataset and running a well-organized shared task at BioNLP 2026.

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

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*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

A Hyperparameter Configuration

Hyperparameter	Value
Base model	`roberta-base`
Parameters	125M
Max sequence length	256 tokens
Dialogue history	$k = 8$ turns
Batch size (train/eval)	16 / 32
Learning rate	2×10^{-5}
LR schedule	Cosine decay
Warmup ratio	0.06
Weight decay	0.01
Gradient clipping	1.0
Max epochs	12
Early stop patience	3 epochs
Early stop metric	Val macro-F1
Focal γ	1.5
Label smooth ε	0.05
Weight dampening	Square root
Hidden dropout	0.1
Attention dropout	0.1
Temperature T	1.3
Precision	fp16
Hardware	NVIDIA T4 (Colab)

Table 4: Full hyperparameter configuration.