AlphaLyrae at SemEval-2026 Task 9: Metric Learning and Asymmetric Loss for Chinese Polarization Analysis

Minh-Hoang Le **Khoan Phung Khac**
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{24520542,24520851}@gm.uit.edu.vn

Abstract

For the Chinese track of SemEval-2026 Task 9 (Detecting Online Polarization), we address two key challenges: polarized content frequently uses implicit language (e.g., homophones and coded terms) to evade moderation, and class distributions exhibit severe long-tail imbalance. We propose a metric learning approach that frames polarization detection as semantic similarity matching, which captures implicit language patterns better than linear decision boundaries. We fine-tune an ERNIE-3.0 encoder with SoftTriple loss and apply kNN retrieval for binary detection (Subtask 1). For multi-label categorization (Subtasks 2 and 3), we transfer learned representations from the detection model and fine-tune with Asymmetric Loss. A priority-based stratified cross-validation strategy ensures minority classes appear across all training folds despite extreme label skew. Evaluated on the official 1,927-sample test set using an end-to-end pipeline, our system achieved Macro-F1 scores of 0.9190 (Rank 6) on Polarization Detection, 0.8244 (Rank 5) on Type Classification, and 0.6670 (Rank 4) on Manifestation Identification.

1 Introduction

Online polarization—the divergence of political and social attitudes to extremes—has become a critical challenge for platform governance, often preceding hate speech and social fragmentation (Naseem et al., 2026b). Unlike simple toxicity, which can frequently be identified via explicit keywords, polarization relies on implicit "us-vs-them" rhetoric, group identity, and stance.

This complexity is magnified in the Chinese language track of SemEval-2026 Task 9. Chinese social media is notorious for high-context communication and rapid linguistic evolution. To evade automated moderation, polarized communities frequently employ linguistic obfuscation—using homophonic puns, acronyms, and cultural markers to signal allegiance without triggering keyword filters. Consequently, standard classification models often struggle to capture the semantic nuances required to detect these implicit divisions.

We propose a metric learning framework for polarization detection. We hypothesize that polarized texts form distinct semantic clusters that are better separated by distance-based metrics than linear boundaries. Our system utilizes an ERNIE-3.0 backbone with a SoftTriple metric learning objective for binary detection (Subtask 1).For the multi-label tasks (Subtasks 2 & 3), we initialize the backbone using weights from the best Subtask 1 model and fine-tune on the polarized subset using Asymmetric Loss (ASL) and class-specific threshold optimization to handle the severe label imbalance.

Our approach achieves strong performance on the complex Chinese dataset. Evaluated on the official test set using a strict end-to-end pipeline (where Subtasks 2 and 3 rely entirely on the predictions of Subtask 1), our system obtained a Macro-F1 of **0.9190** (Rank 6) on Polarization Detection, **0.8244** (Rank 5) on Type Classification, and **0.6670** (Rank 4) on Manifestation Identification. We release our code to facilitate future research on multilingual polarization.[1]

2 Background

2.1 Task Definition

SemEval-2026 Task 9 (POLAR) (Naseem et al., 2026a) establishes a hierarchical pipeline for analyzing online polarization. We participate in the Chinese track, which consists of three cascaded subtasks:

1. **Subtask 1 (Detection):** Binary classification identifying whether a text exhibits polarized rhetoric.

[1] https://github.com/lmhoang06/AlphaLyrae-SemEval-2026-Task-9

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 503–508
July 3-4, 2026

2. **Subtask 2 (Type):** Multi-label classification of the polarization topic (e.g., *Gender/Sexual*, *Political*, *Racial/Ethnic*).

3. **Subtask 3 (Manifestation):** Multi-label classification of the rhetorical strategy (e.g., *Vilification*, *Dehumanization*).

The specific difficulty of the Chinese track is exemplified by training sample `zho_715...476`: "女权=女拳=汉奸=敌人" (lit. "Feminism = 'Fem-fists' [homophone] = Traitors = Enemy"). Here, the author uses a homophone (*quan* for 'fist' rather than 'rights') to implicitly equate gender identity with political subversion. Detecting this requires the model to simultaneously recognize *Gender* and *Political* topics (Subtask 2) while identifying *Vilification* and *Extreme Language* (Subtask 3).

2.2 Dataset Statistics

The dataset aggregates content from Chinese platforms including Weibo, Zhihu, and Tieba. The organizers provided a training set of 4,280 samples, a development set of 214 samples, and a held-out test set of 1,927 samples.

Class Imbalance The Chinese subset exhibits severe distributional challenges. While Subtask 1 is balanced (49.5% polarized), the downstream multi-label tasks suffer from extreme long-tail imbalance within the 2,121 polarized training samples. In Subtask 2, *Racial/Ethnic* polarization dominates (45.6%), whereas *Religious* polarization is exceedingly rare (4.0%). Similarly, in Subtask 3, *Stereotype* accounts for over 60% of cases, while *Invalidation* accounts for under 10%. This extreme skew necessitates specialized cross-validation and Asymmetric Loss strategies detailed in Section 4.

3 Methodology

3.1 Shared Model and Training Strategy

Backbone Encoder We adopt ERNIE-3.0 (Sun et al., 2021) as our shared encoder. ERNIE-3.0 employs knowledge-enhanced pre-training, which improves the representation of Chinese entities and concepts. We extract the `[CLS]` embedding from the final layer as the global sentence representation, which serves as input to the task-specific methods (§3.2 and §3.3).

Layer-wise Learning Rate Decay (LLRD) To stabilize fine-tuning, we apply LLRD (Howard and Ruder, 2018) where $\eta_l = \eta_{head} \cdot \xi^{L-l}$. This assigns higher learning rates to upper layers, allowing the classification head to adapt quickly while preserving representations in lower layers.

Adversarial Training We employ the Fast Gradient Method (FGM) (Miyato et al., 2017) to improve robustness against label noise. We perturb input embeddings by $r_{adv} = \epsilon \cdot g/\|g\|_2$, where $g = \nabla_x \mathcal{L}(f(x;\theta), y)$, to promote smoother decision boundaries.

K-Fold Ensemble Strategy We use 5-fold cross-validation across all subtasks. We train $K = 5$ independent models, one per fold. At inference, we average the probability outputs from all K models to reduce prediction variance.

3.2 Subtask 1: Polarization Detection via Metric Learning

We frame polarization detection as a metric learning task. Metric learning groups semantically similar texts by distance, providing more flexible decision boundaries than linear classification for diverse polarized topics.

Metric Projection & Loss We use a projection head (Linear $\rightarrow$ GELU $\rightarrow$ LayerNorm $\rightarrow$ Linear) to map the `[CLS]` representation to $\mathbb{R}^d$ (see §4). The output z is L_2-normalized to the unit hypersphere. We fine-tune the model end-to-end using SoftTriple loss (Qian et al., 2019). SoftTriple maintains multiple centers per class to represent the diversity of topics within the "Polarized" and "Neutral" categories. We select checkpoints based on Precision@1 (P@1) on the validation set, the standard metric for evaluating metric learning models.

Ensemble Retrieval Inference We use the shared K-fold ensemble (§3.1) with kNN retrieval. For a test sample x, each fold model f projects x and retrieves the k nearest neighbors $\mathcal{N}_f(x)$ from its corresponding training fold. The final polarization probability is the average vote across all folds:

$$\hat{y}(x) = \frac{1}{K} \sum_{f=1}^{K} \frac{1}{k} \sum_{i \in \mathcal{N}_f(x)} y_i \qquad (1)$$

where $y_i \in \{0, 1\}$ is the label of neighbor i. We apply a classification threshold τ to obtain binary predictions: $\text{pred}(x) = \mathbb{1}[\hat{y}(x) \geq \tau]$. The values of k and τ are specified in §4.

3.3 Subtasks 2 & 3: Multi-Label Classification

We formulate Polarization Type (Subtask 2) and Manifestation (Subtask 3) as multi-label classification tasks. We use the shared strategies (K-Fold ensemble, LLRD, FGM) described in §3.1.

Transfer Learning & Architecture. We initialize the backbone with weights from the best-performing Subtask 1 fold model (selected by validation P@1). This transfers learned representations to the multi-label tasks. The classification head consists of: Linear → LayerNorm → GELU → Dropout → Linear.

Asymmetric Loss. We train using Asymmetric Loss (ASL) (Ben-Baruch et al., 2020), which extends Focal Loss with separate focusing parameters for positive and negative samples. ASL down-weights easy negatives to focus on hard examples and positive classes, addressing the label imbalance in multi-label settings.

Threshold Optimization. We tune class-specific thresholds to balance precision and recall. For each class, we select the threshold minimizing $|P - R|$ (with F1 as tie-breaker); see §4 for grid search details.

4 Experimental Setup

4.1 Evaluation Protocol

Data Splitting We evaluate all models using Stratified 5-Fold Cross-Validation. We use a fixed random seed (42) for all stochastic operations (fold splitting, model initialization, and batch shuffling) to ensure strict reproducibility.

For **Subtask 1**, stratifying solely on binary labels fails to preserve the distribution of specific polarization types across folds. We construct single-label proxy targets from Subtask 2 labels using a priority hierarchy: *Gender/Sexual* (34.1%) > *Religious* (4.0%) > *Racial/Ethnic* (45.6%) > *Political* (11.8%) > *Other* (17.4%).

This hierarchy captures distinct polarization modes. We prioritize *Gender/Sexual* first as it forms a distinct semantic cluster. We prioritize *Religious* next to prevent fold starvation of the rarest minority class. Finally, we prioritize *Racial/Ethnic* over *Political* to group frequent overlaps (e.g., nationalist rhetoric) under the dominant Racial category, reserving the Political target for distinct ideological content. Samples with multiple labels are assigned to the highest-priority category.

For **Subtasks 2 & 3**, we filter the dataset to the 2,121 polarized samples and apply Multilabel Stratified K-Fold directly on the target labels.

Metrics For Subtask 1, we use **Precision@1** (P@1) as the primary metric for early stopping. We specifically chose P@1 over Macro-F1 for validation because it directly aligns with our metric learning objective, which optimizes distance-based retrieval rather than linear decision boundaries. For Subtasks 2 & 3, we use **Macro-F1**. All metrics for early stopping and threshold optimization are computed on the validation fold of each cross-validation split.

4.2 Implementation Details

All experiments were conducted on a single **NVIDIA RTX 4060 Laptop GPU (8GB)** using PyTorch and `bf16` mixed precision. All reported results were obtained on this configuration to demonstrate the feasibility of the method on consumer-grade hardware.

Shared Configuration We use `nghuyong/ernie-3.0-xbase-zh`[2] (12 layers, 1024 hidden size) as the backbone. We set the maximum sequence length to 96 tokens, which covers 100% of the training and test samples without truncation. Models are trained for a maximum of 3,000 steps per fold with evaluation every 100 steps. We use Early Stopping with a patience of 5 evaluations (500 steps) based on the validation fold performance.

We optimize using AdamW (Loshchilov and Hutter, 2017) with a weight decay of 0.01 and a warmup ratio of 0.1. The peak learning rate is set to 2×10^{-5} for the classification head and the top encoder layer (Layer 12). To preserve pre-trained knowledge, we apply **Layer-wise Learning Rate Decay (LLRD)** with $\xi = 0.95$, where the learning rate for layer l is $\text{LR}_l = 2 \times 10^{-5} \cdot 0.95^{12-l}$. **Adversarial Training (FGM)** with perturbation $\epsilon = 0.2$ is also applied to improve robustness.

4.3 Task-Specific Configuration

Subtask 1 Settings We train with a batch size of 24. The metric projection head consists of two linear layers ($1024 \rightarrow 1024 \rightarrow 256$) to map embeddings to a 256-dimensional unit hypersphere.

[2] A PyTorch-converted version of the official Baidu ERNIE 3.0 model, available on HuggingFace.

For our official submission, we set $k = 49$ for kNN retrieval and applied a classification threshold of $\tau = 0.4$. These hyperparameters were selected empirically based on preliminary observations during development, prior to systematic optimization. Post-hoc ablation analysis (§5) explores the impact of these choices.

Subtasks 2 & 3 Settings We use a batch size of 8 to accommodate the 8GB GPU memory constraints. We apply a dropout rate of 0.1. The classification head projects $1024 \rightarrow 1024 \rightarrow N_{classes}$. We optimize using Asymmetric Loss (ASL) with $\gamma_+ = 1.0$ and class-specific γ_- values tuned on validation data:

- **Subtask 2:** $\gamma_- = [4.0, 3.0, 3.0, 4.0, 2.0]$ for *Political*, *Racial*, *Religious*, *Gender*, and *Other*, respectively.
- **Subtask 3:** $\gamma_- = [4.0, 4.0, 3.5, 3.5, 2.0, 2.0]$ for *Stereotype*, *Vilification*, *Dehumanization*, *Extreme Lang.*, *Lack of Empathy*, and *Invalidation*.

For threshold optimization (described in §3.3), we perform a grid search over $[0.1, 0.9]$ with a step size of 0.01 on the validation fold.

5 Results and Analysis

5.1 Main Results

Our system achieved Macro-F1 scores of 0.9190 (Rank 6) on Polarization Detection, 0.8244 (Rank 5) on Type Classification, and 0.6670 (Rank 4) on Manifestation Identification on the official hidden test set. These results were obtained using an end-to-end pipeline where Subtasks 2 and 3 rely entirely on Subtask 1 predictions.

5.2 Ablation Study

To evaluate our design choices, we conduct ablation studies on the development set (Table 1). To measure the upper-bound performance of each configuration, we report Macro-F1 scores using *oracle thresholds*. Specifically, classification thresholds were optimized directly on the development set via grid search. While this introduces threshold data leakage, it isolates the model's representation learning from variance in threshold selection. Subtasks 2 and 3 ablations evaluate only the polarized subset (gold labels) to isolate multi-label performance from binary detection errors. Note that these oracle scores represent upper bounds on the development set and do not reflect expected test set generalization.

Handling Extreme Imbalance. Asymmetric Loss (ASL) provided the largest performance improvement in our ablations. Replacing ASL with standard Binary Cross Entropy (BCE) dropped the Subtask 2 score by 5.79 F1 points ($0.8312 \rightarrow 0.7733$), demonstrating that dynamically down-weighting easy negative samples is essential for capturing long-tail categories like *Religious* polarization.

Negative Transfer in Hierarchical Pipelines. Because Subtasks 2 and 3 train on a subset of Subtask 1 data, the encoder has already seen these texts during binary detection training. We expected that initializing the model with Subtask 1 weights would improve downstream performance. However, we observed minimal differences: without initialization, Subtask 2 achieved 0.8316 (vs. 0.8312) and Subtask 3 achieved 0.7448 (vs. 0.7422). Both differences ($\Delta < 0.003$) are well within margin of noise.

More surprisingly, the completely unregularized baseline outperformed our final system on Subtask 3 (0.7642 vs. 0.7422). This suggests that the metric learning objective in Subtask 1 may create embedding spaces that are less suitable for fine-grained multi-label classification, though controlled experiments are needed to confirm this hypothesis.

Regularization Trade-offs. FGM and LLRD showed contrasting effects across the tasks. In Subtask 1, removing only FGM slightly improved the development F1 to 0.9437, suggesting slight overfitting prevention comes at a small performance cost. In contrast, Subtask 3 performance dropped sharply without FGM ($0.7422 \rightarrow 0.6470$), but removing *both* FGM and LLRD recovered much of this loss (0.7282). This pattern indicates that FGM is particularly important for Subtask 3 when combined with LLRD's conservative learning rates, possibly by helping the model adapt from the Subtask 1 initialization to the multi-label objective.

6 Limitations

While our system demonstrates strong empirical performance, several limitations provide avenues for future work. First, the ablation study utilizes oracle thresholds tuned on the development set to isolate representation learning; performance using

Model Configuration	Subtask 1 (Dev F1)	Subtask 2 (Dev F1)	Subtask 3 (Dev F1)
Final System	0.9392	0.8312	0.7422
– Task-Specific Components –			
w/o Asymmetric Loss (uses BCE)	-	0.7733	0.7365
w/o Subtask 1 Initialization	-	**0.8316**	0.7448
– Shared Regularization –			
w/o FGM	**0.9437**	0.8274	0.6470
w/o FGM & w/o LLRD	0.9393	0.8034	0.7282
Unregularized Baseline*	0.9346	0.7839	**0.7642**

Table 1: Ablation study on the development set with oracle thresholds (tuned directly on the evaluation set). Subtasks 2 and 3 evaluate only the polarized subset. Bold indicates the highest dev score per subtask. *The Unregularized Baseline for Subtask 1 uses a standard linear classification head (no metric learning), no FGM, and no LLRD. For Subtasks 2 and 3, it removes Subtask 1 initialization, ASL (uses BCE), FGM, and LLRD.

thresholds tuned strictly on training folds may exhibit slight degradation in real-world generalization. Second, while our metric learning approach was designed to capture implicit language (e.g., homophones), conducting a granular robustness evaluation separating direct insults from obfuscated slang requires a specialized, human-annotated subcorpus, which we leave for future work. Finally, our ablations suggest a potential representation conflict when transferring weights from binary detection to multi-label classification. Future work could investigate decoupling representation learning across stages—such as utilizing adapter layers or freezing the backbone—to mitigate this negative transfer.

7 Conclusion

We present a metric learning framework for the Chinese track of SemEval-2026 Task 9, addressing the challenges of implicit language and severe class imbalance. On the official test set, our end-to-end system achieved competitive rankings: 6th for Polarization Detection (0.9190), 5th for Type Classification (0.8244), and 4th for Manifestation Identification (0.6670). These results were achieved entirely on consumer-grade hardware (an 8GB RTX 4060), demonstrating that effective polarization detection is feasible on consumer hardware. Our ablation studies underscore that Asymmetric Loss is critical for capturing rare long-tail categories, and that standard cross-entropy baselines fail to properly regularize the severe label skew inherent in Chinese social media discourse. They also reveal a representation conflict when transferring metric learning embeddings to fine-grained multi-label tasks, suggesting that decoupled training strategies merit future investigation.

Acknowledgments

We would like to thank Ngan Luu-Thuy Nguyen for facilitating the collaboration with our mentor. We are also grateful to Duc-Vu Nguyen for his guidance and insightful feedback during the development of this project.

References

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2020. Asymmetric loss for multi-label classification. *Preprint*, arXiv:2009.14119.

Jeremy Howard and Sebastian Ruder. 2018. 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)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. 2019. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6450–6458.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, and 3 others. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *Preprint*, arXiv:2107.02137.