IReL_IIT(BHU) at SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization

Soumadip Majumder[1] **Arjun Mukherjee**[1] **Krishna Tewari**[1]
Sanjaya Kumar Lenka[2] **Sukomal Pal**[1]
[1]Department of Computer Science and Engineering
[2]Department of Humanistic Studies
Indian Institute of Technology (BHU) Varanasi
Varanasi, India
{soumadip.majumder.cse23, arjunmukherjee.rs.cse23, krishnatewari.rs.cse24,
slenka.hss, spal.cse}@iitbhu.ac.in

Abstract

This paper presents the IReL_IIT(BHU) submission to SemEval-2026 Task 9 for the Chinese language track. We participated in all three subtasks: binary polarization detection, multi-label polarization type classification, and multi-label manifestation identification. Our approach is based on a unified transformer-based framework with cross-validation, prediction aggregation, and threshold optimization to improve robustness across tasks. On the official evaluation, our systems achieved Macro-F1 scores of 0.9081, 0.7962, and 0.6484 for Subtasks 1, 2, and 3, respectively on test data.

1 Introduction and Related Work

Online polarization poses a significant challenge for natural language processing, as it shapes public discourse, intensifies inter-group conflict, and amplifies harmful narratives. Recent benchmarks such as SemEval-2026 Task 9 move beyond binary detection to require fine-grained modeling of polarization, including its targets and manifestations across multilingual and multicultural settings (Naseem et al., 2026a,b).

Recent advances further emphasize multilingual, multi-dimensional, and context-aware formulations of polarization, as reflected in large-scale benchmarks such as POLAR (Naseem et al., 2025) and extensions to low-resource and real-world settings (Davoudi and Goharian, 2026; Sermpezis et al., 2026). Concurrently, studies on real-world geopolitical discourse highlight the dynamic and evolving nature of polarization across events and contexts (Sermpezis et al., 2026). Moreover, recent analyses show that large language models often inherit and amplify political biases (Feng et al., 2023), motivating the need for robust, calibrated, and evaluation-aware approaches for reliable polarization detection (Rim et al., 2026).

Transformer-based models have become the standard for such tasks due to their strong contextual representations (Devlin et al., 2019; Liu et al., 2019). In Chinese NLP, pretrained models such as MacBERT (Cui et al., 2020) and Erlangshen-DeBERTa (Zhang et al., 2022; He et al., 2021) provide robust foundations. However, existing approaches largely adopt task-specific pipelines, treating binary detection and multi-label classification independently. In multi-label settings, standard sigmoid-based formulations (Zhang and Zhou, 2014) with class weighting (Lin et al., 2017) are commonly used, but they often rely on fixed thresholds and lack systematic calibration. Similarly, while cross-validation and ensembling are widely applied to improve robustness (Kohavi, 1995; Dietterich, 2000), they are typically used in an ad-hoc manner, without fully exploiting out-of-fold (OOF) predictions for unified calibration and model combination (Niculescu-Mizil and Caruana, 2005; Guo et al., 2017). As a result, prior systems often lack consistency, reproducibility, and principled optimization across subtasks. We propose a unified transformer framework leveraging cross-validation, OOF-based calibration, and ensembling for robust, reproducible multi-dimensional polarization detection.

2 Data and Preprocessing

We use the official Chinese training and test sets provided by the shared task (Naseem et al., 2026b). The training split contains input texts with gold labels, while the test split includes only text and instance identifiers.

We apply minimal preprocessing to preserve the original text, including handling missing values, casting inputs to string, collapsing repeated whitespace, and removing corrupted characters. We avoid aggressive normalization, as transformer tokenizers already capture surface-level variation and excessive preprocessing may discard useful signals for polarization detection.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2354–2358

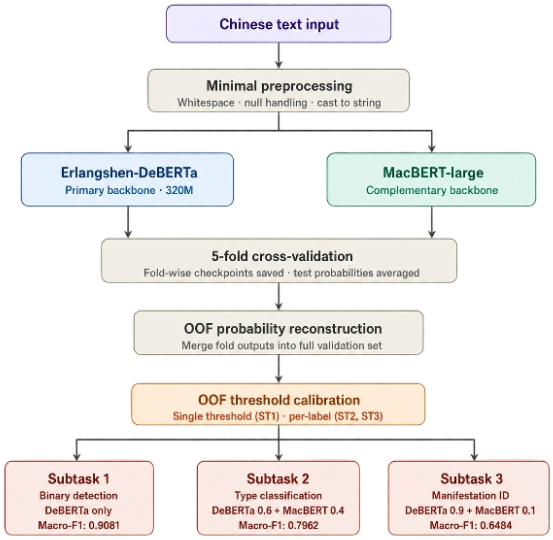

Figure 1: Architectural Pipeline

3 System Overview

Although the three subtasks differ in their output structure, all systems follow a unified workflow as shown in Figure 1. We fine-tune pretrained Chinese transformer encoders using stratified 5-fold cross-validation and collect out-of-fold (OOF) probabilities for validation samples. These OOF predictions serve as a common basis for both threshold calibration and ensemble weight selection. At inference time, fold-wise predictions are averaged, optionally combined via weighted ensembling, and converted to final labels using tuned decision thresholds.

3.1 Backbone Models

We experiment with two pretrained Chinese encoders: IDEA-CCNL/Erlangshen-DeBERTa-v2-320M-Chinese (Zhang et al., 2022) and hfl/chinese-macbert-large (Cui et al., 2020). Erlangshen-DeBERTa consistently provides the strongest single-model performance, while MacBERT offers complementary behavior, particularly in multi-label settings.

3.2 Unified Training and Calibration Strategy

All experiments use 5-fold cross-validation. For each fold, the model is trained on four partitions and validated on the remaining split. Validation probabilities and indices are stored and merged to reconstruct full OOF predictions.

These OOF predictions are used for: (i) threshold tuning, and (ii) ensemble weight selection.

Instead of using a fixed threshold (e.g., 0.5), we directly optimize decision thresholds on OOF predictions to maximize Macro-F1. For multi-model systems, we perform a grid search over candidate ensemble weights, followed by threshold tuning on the resulting OOF probabilities as shown in Table 1. The best configuration is selected based on OOF Macro-F1.

Table 1: Summary of thresholding and ensemble strategies across subtasks.

Subtask	Thresholding	Ensemble Strategy
Subtask 1	Single threshold	Single model (DeBERTa)
Subtask 2	Per-label thresholds	0.6 / 0.4 (DeBERTa / MacBERT)
Subtask 3	Per-label thresholds	0.9 / 0.1 (DeBERTa / MacBERT)

3.3 Subtask-Specific Modeling

This section describes the specific settings for each subtask.

Subtask 1: Binary Polarization Detection We formulate this task as binary sequence classification and fine-tune Erlangshen-DeBERTa-v2-320M-Chinese with a 2-way classification head using stratified 5-fold cross-validation. Although MacBERT and simple ensembling slightly improve OOF performance, they do not generalize as well on the hidden test set. Therefore, we adopt a single DeBERTa model with fold-wise probability averaging and an OOF-tuned threshold (slightly below 0.5).

Subtask 2: Multi-Label Polarization Type Classification This task is modeled as a 5-label multi-label classification problem using a sigmoid output layer and `BCEWithLogitsLoss` with per-label `pos_weight` to address class imbalance. We train both DeBERTa and MacBERT under the same setup.

Final predictions are obtained via global weighted probability averaging, followed by per-label threshold tuning on OOF predictions. A DeBERTa-dominant ensemble (0.6/0.4) achieves the best performance, indicating complementary modeling of polarization types.

Subtask 3: Multi-Label Manifestation Identification For this task, we use a sigmoid-based multi-label formulation over six manifestation categories, trained with a BCE-based objective and per-label `pos_weight`.

We explore both global and per-label ensemble weighting strategies. Although per-label weighting

improves OOF performance, it does not generalize well. A DeBERTa-dominant global ensemble (0.9/0.1) provides more stable results, suggesting higher sensitivity of this task to overfitting and train-test mismatch.

4 Experimental Setup

This section describes the implementation details, training configurations, and evaluation protocol used to assess the proposed unified framework across all subtasks.

4.1 Implementation and Evaluation Protocol

All experiments were implemented in Python using PyTorch and Hugging Face Transformers. Training was conducted in a Conda environment on an NVIDIA GeForce RTX 4060 Laptop GPU with FP16 mixed-precision when supported. To ensure consistency across subtasks, we employed a unified training strategy based on stratified 5-fold cross-validation with seed 42. Common hyperparameters are summarized in Table 2.

We participated in the Chinese (zho) track for all subtasks, using the official training split for development and generating predictions on the hidden test set without altering the provided data partitions. Performance was evaluated using Macro-F1, which also served as the model selection criterion.

Table 2: Shared training configuration across subtasks.

Component	Setting
Cross-validation	5-fold (stratified)
Random seed	42
Optimizer	AdamW
Learning rate	1.5e-5 (DeBERTa), 1.5e-5–1.8e-5 (MacBERT)
Weight decay	0.01
Warmup ratio	0.06
Dropout	0.1
Precision	FP16
Loss (multi-label)	BCEWithLogitsLoss
Class imbalance handling	Per-label `pos_weight`
Early stopping	Patience = 2
Inference aggregation	Fold-wise probability averaging
Calibration	OOF-based threshold tuning

4.2 Subtask-Specific Settings

Although the overall pipeline is shared, subtasks differ in sequence length, batch size, and modeling strategy.

Subtask 1: Binary Polarization Detection We evaluated multiple pretrained encoders, including Erlangshen-DeBERTa-v2-320M-Chinese, MacBERT, mDeBERTa-v3, and XLM-R. The final system uses Erlangshen-DeBERTa-v2-320M as a single-model solution.

Training was performed with a maximum sequence length of 128 and batch size 16. Final predictions were obtained by averaging fold-wise probabilities, followed by tuning a single decision threshold on OOF predictions to maximize Macro-F1.

Subtask 2: Polarization Type Classification This task is formulated as multi-label classification with a sigmoid output layer. We trained both Erlangshen-DeBERTa-v2-320M and MacBERT using a maximum sequence length of 192.

Final predictions were obtained via weighted probability averaging (DeBERTa/MacBERT), followed by per-label threshold tuning on OOF predictions.

Subtask 3: Manifestation Identification Similar to Subtask 2, this task uses a multi-label formulation with sigmoid outputs and BCE-based optimization. Models were trained with a longer sequence length of 256.

We explored both global and per-label ensemble weighting strategies. While per-label weighting improved validation performance, global weighted averaging generalized better to the test set. Final predictions were obtained using OOF-tuned per-label thresholds.

5 Results

This section presents the experimental results of our approach across all subtasks.

5.1 Development-Phase Results

Table 3 summarizes the performance of our best systems on the visible development phase of Codabench for the Chinese track.

Table 3: Development-phase performance on the visible Codabench split (Chinese track).

Subtask	Model	Macro-F1
Subtask 1	DeBERTa	0.9205
Subtask 2	DeBERTa + MacBERT (0.6 / 0.4)	0.8110
Subtask 3	DeBERTa + MacBERT (0.9 / 0.1)	0.7136

As shown in Table 3, a consistent pattern emerges across tasks. For Subtask 1, the single DeBERTa model achieves the best performance, indicating that binary polarization detection benefits from a strong and well-calibrated model without requiring ensembling. Although ensembling slightly improves the OOF proxy, it does not translate into better generalization.

In contrast, Subtask 2 benefits from weighted ensembling, with the DeBERTa-dominant (0.6/0.4) configuration achieving the best performance. This suggests that different models capture complementary aspects of polarization targets. Label-wise analysis shows strong performance on religious, gender/sexual, and racial/ethnic categories, while the *other* category remains the most challenging, likely due to its residual and overlapping nature.

For Subtask 3, a more DeBERTa-dominant ensemble (0.9/0.1) performs best, while more flexible strategies such as per-label weighting improve OOF scores but do not generalize well. This indicates a higher sensitivity to overfitting and train–test mismatch. Among the labels, vilification, dehumanization, and stereotype achieve higher scores, whereas lack of empathy remains the most difficult category.

We further observe that threshold tuning based on OOF predictions consistently improves performance, particularly in multi-label settings where fixed thresholds are suboptimal under Macro-F1 evaluation.

Overall, tasks relying on explicit lexical cues (Subtasks 1 and 2) are more stable, while manifestation identification (Subtask 3) remains challenging due to its reliance on implicit and context-dependent signals.

5.2 Test-Phase Results

Table 4 summarizes the official results on the hidden test set. Our approach achieves strong performance across all subtasks, with the most stable results in binary polarization detection, while multi-label tasks benefit from ensembling.

Table 4: Official test-phase results on the hidden test set (Chinese track).

Subtask	Macro-F1	Rank
Subtask 1	0.9081	10th
Subtask 2	0.7962	8th
Subtask 3	0.6484	7th

Subtask 3 exhibits a clear generalization gap ($0.7136 \rightarrow 0.6484$), highlighting the difficulty of modeling manifestation categories. Label-wise trends indicate that stereotype remains relatively stable, whereas dehumanization and invalidation degrade more noticeably, and lack of empathy remains the most challenging class. This suggests that manifestation labels rely more on implicit and context-dependent cues.

We also observe that OOF-based threshold tuning contributes to consistent performance, particularly in multi-label settings under Macro-F1 evaluation. Overall, while well-calibrated transformer models provide strong baselines, tasks requiring implicit semantic understanding remain challenging.

6 Conclusion

We presented a unified transformer-based framework for SemEval-2026 Task 9 (Chinese track), addressing binary polarization detection, multi-label type classification, and manifestation identification. Our approach integrates cross-validation, OOF-based probability reconstruction, threshold calibration, and lightweight ensembling within a single pipeline.

The results demonstrate strong performance across all subtasks, with a single well-calibrated model sufficient for binary detection, while multi-label tasks benefit from ensembling. However, manifestation identification remains challenging due to its reliance on implicit and context-dependent cues, leading to a noticeable generalization gap.

Overall, our findings highlight the importance of calibration and disciplined validation design, and suggest that future work should focus on better modeling implicit semantics in polarization.

7 Limitations and Ethical Considerations

Our work uses standard pretrained transformer models and a unified pipeline without explicitly modeling dependencies between labels, which may limit its ability to capture structured polarization patterns. While OOF-based calibration improves performance, threshold tuning may be sensitive to validation data and may not generalize under distribution shifts, particularly for manifestation identification. The system operates on potentially sensitive content (e.g., political, religious, and social topics), and misclassification could affect downstream moderation or analysis. As with most data-driven approaches, the model may inherit biases present in the training data, including cultural or linguistic biases in the Chinese track. We do not release any new dataset and use only publicly provided shared-task data. No personally identifiable information is introduced by our pipeline.

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online.

Saeedeh Davoudi and Nazli Goharian. 2026. Online polarization detection in Persian (Farsi) social media. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 50–59, Rabat, Morocco. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems (MCS)*, pages 1–15. Springer.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327.

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*, abs/1907.11692.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, 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.

Usman Naseem, Juan Ren, Saba Anwar, Sarah Kohail, Rudy Alexandro Garrido Veliz, Robert Geislinger, Aisha Jabr, Idris Abdulmumin, Laiba Qureshi, Aarushi Ajay Borkar, Maryam Ibrahim Mukhtar, Abinew Ali Ayele, Ibrahim Said Ahmad, Adem Ali, Martin Semmann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632.

Nakwon Rim, Joshua Conrad Jackson, Marc G Berman, and Yuan Chang Leong. 2026. Natural language reveals that political partisans are more affectively aligned over political issues than partisan identities. *Communications Psychology*, 4(1):65.

Pavlos Sermpezis, Stelios Karamanidis, Eva Paraschou, Ilias Dimitriadis, Sofia Yfantidou, Filitsa-Ioanna Kouskouveli, Thanasis Troboukis, Kelly Kiki, Antonis Galanopoulos, and Athena Vakali. 2026. Agoraspeech: a multi-annotated comprehensive dataset of political discourse through the lens of humans and AI. *Journal of Computational Social Science*, 9(2):36.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, and 6 others. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.