SemEval-2026 Task 9: A Two-Stage Framework for Tackling Class Imbalance in Multilingual Polarization Detection

Mohammed Nagi
African Institute for Mathematical Sciences (AIMS), South Africa
esameldin@aims.ac.za

Abstract

This paper describes our system submission for SemEval-2026 Task 9: Detecting Multilingual, Multicultural, and Multievent Online Polarization. We address all three subtasks, Polarization Detection, Type Classification, and Manifestation Identification, for both English and Arabic. The primary challenge in this task is the extreme class imbalance (up to 16:1) and the complexity of multi-label classification in low-resource settings. To mitigate this, we propose a robust two-stage training framework. Our approach leverages language-specific pre-trained transformer backbones (**DeBERTa-v3-base** for English and **MARBERTv2** for Arabic) optimised with Focal Loss to handle distributional skew. Furthermore, we implement a dynamic, per-class thresholding strategy derived from validation data to maximize the Macro-F1 score. Our experimental results demonstrate that this targeted approach, which prioritises data efficiency and loss-function engineering over model scaling, yields competitive performance and significant improvements in minority class recall.

1 Introduction

The proliferation of polarized discourse on social media platforms poses a significant threat to social cohesion and democratic dialogue. SemEval-2026 Task 9 addresses this critical issue by formulating the problem of polarization detection across three granularities: binary detection (Subtask 1), categorization of the target group (Subtask 2), and identification of the rhetorical manifestation (Subtask 3).

While recent advances in Large Language Models (LLMs) have shown promise in varied NLP tasks, detecting polarization requires a nuanced understanding of cultural context, slang, and implicit biases often found in short-text social media posts. Furthermore, the task presents a significant distributional challenge: the datasets are characterized by severe class imbalance. For instance, political polarization is abundantly represented, while categories such as gender or religious polarization are exceedingly rare. Standard optimization objectives, such as Cross-Entropy Loss, often fail in such scenarios, as the model achieves high accuracy by simply predicting the majority class and ignoring the critical minority instances.

Our system is designed to tackle these specific challenges. We prioritize a data-centric approach over parameter scaling. Instead of utilizing massive, computationally expensive models, we focus on: (1) selecting domain-adapted pre-trained models that align with the linguistic characteristics of the data (e.g., Arabic dialects); (2) employing Focal Loss to dynamically down-weight the contribution of easy samples and focus learning on hard, minority class examples; and (3) implementing a two-stage training pipeline that decouples threshold optimisation from model fine-tuning. This paper details our methodology, provides a comprehensive exploratory data analysis (EDA) that informed our design choices, and presents an ablation study validating our contributions.

2 Exploratory Data Analysis

A rigorous statistical analysis of the provided training data was conducted to inform our system architecture and hyperparameter selection. The dataset comprises approximately 3,000 to 3,500 annotated instances per language. Table 1 summarizes the key statistics.

2.1 Class Distribution Imbalance

The most pervasive challenge identified is the long-tail distribution of classes. In English Subtask 2, the "Political" category accounts for 35.7% of the data, whereas "Gender/Sexual" polarization constitutes only 2.2%, resulting in a staggering 16:1 imbalance ratio. While the Arabic dataset is more

Task	Lang	N	Avg Len	Pos%	Imbal.
S1	ENG	3222	12.3	36.5	-
	ARB	3380	16.7	44.7	-
S2	ENG	3222	12.3	-	16.0:1
	ARB	3380	16.7	-	2.8:1
S3	ENG	3222	12.3	-	2.4:1
	ARB	3380	16.7	-	4.6:1

Table 1: Dataset statistics. N = sample count, Avg Len = average word count, Pos% = positive class percentage (Subtask 1), Imbal. = imbalance ratio between most and least frequent classes (Subtasks 2 and 3).

balanced (2.8:1), the skew is still significant enough to hinder the learning of minority classes. This observation directly motivated our adoption of Focal Loss, which is specifically designed to prevent the vast number of easy negatives (majority class examples) from overwhelming the gradient during training.

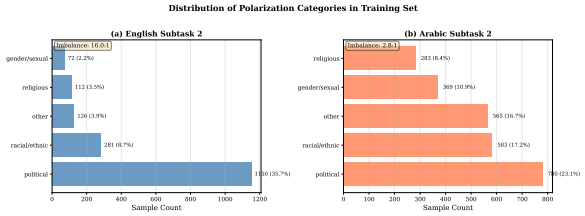

Figure 1: Distribution of Polarization Categories in the Training Set for Subtask 2. Note the severe imbalance in English (left) compared to the relatively smoother distribution in Arabic (right).

2.2 Multi-Label Complexity and Co-occurrence

Subtasks 2 and 3 are multi-label classification problems, meaning a single text can exhibit multiple types of polarization or manifestations simultaneously. Our analysis revealed that a significant portion of the data (13.2% for English S2 and 39.0% for Arabic S3) contains multiple labels.

Furthermore, we analyzed label co-occurrence to understand the dependencies between categories. Figure 2 illustrates the correlation matrix for Arabic Subtask 3. We observed strong positive correlations ($r > 0.6$) between *Vilification*, *Dehumanization*, and *Extreme Language*. This suggests that these rhetorical devices are often employed together in polarized discourse. Conversely, *Lack of Empathy* showed weaker correlations ($r < 0.3$), indicating it is a distinct, perhaps more subtle, form of polarization that may require the model to capture deeper semantic nuances rather than explicit lexical triggers.

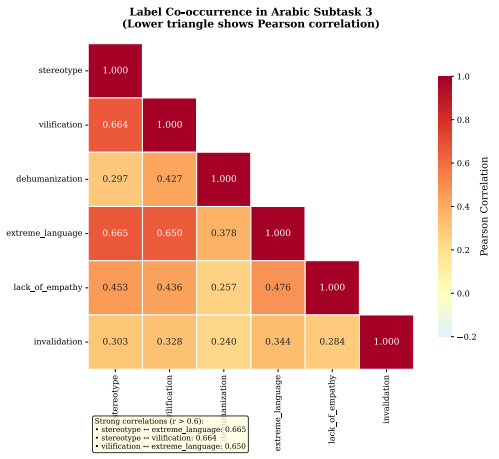

Figure 2: Label co-occurrence correlation matrix for Arabic Subtask 3. Strong correlations indicate that manifestations like Vilification and Dehumanization frequently co-occur.

2.3 Text Length Characteristics

We analysed the distribution of text lengths to determine an appropriate maximum sequence length for our transformer models. As shown in Figure 3, Arabic texts are on average longer than English texts (16.7 vs. 12.3 words). However, the 95th percentile for both languages falls below 35 words. Based on this, we selected a maximum sequence length of 128 tokens. This choice provides a safe buffer for sub-word tokenization while maintaining computational efficiency and minimizing padding.

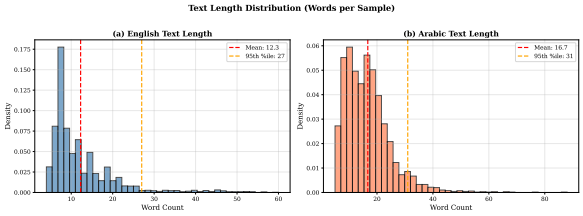

Figure 3: Text length distribution (word count). While Arabic texts are longer on average, the vast majority of samples in both languages fit comfortably within standard transformer context windows.

3 Methodology

Our proposed system is a robust, data-efficient pipeline designed to handle the specific challenges identified in our EDA.

3.1 Preprocessing

We applied standard text cleaning procedures to reduce noise. This included the normalization of URLs and user mentions to placeholder tokens (e.g., `[URL]`, `@USER`) and the removal of non-standard whitespace characters. For Arabic, we performed specific text normalization steps, such as unifying different forms of Alef and normalizing Ta-Marbuta to Ha, to align the input text with the pre-training data of the MARBERT model.

3.2 Model Architecture

We eschewed generic multilingual models (like mBERT or XLM-R) in favor of language-specific experts. Our experiments indicated that domain-specific pre-training is crucial for capturing the nuances of social media discourse.

English: We adopted a hybrid strategy. For Subtask 2 (Topic Classification), we utilized **DeBERTa-v3-base**, leveraging its disentangled attention mechanism which is highly effective for semantic content classification. However, for Subtasks 1 and 3 (Polarization Detection and Manifestation Identification), we employed **cardiffnlp/twitter-RoBERTa-base-sentiment-latest** (Loureiro et al., 2022). Since polarization manifestations (such as *Vilification* and *Lack of Empathy*) are intrinsically linked to emotional tone and social media vernacular (hashtags and slang), this model, which was pre-trained on 124M tweets, demonstrated superior performance in detecting these subtle signals.

Arabic: We utilized **MARBERTv2** (Abdul-Mageed et al., 2021). Unlike standard AraBERT models trained on Modern Standard Arabic (MSA) news corpora, polarization on social media is frequently expressed in dialectal Arabic. MARBERTv2 is pre-trained on a massive corpus of 1 billion Arabic tweets, providing superior coverage of the informal lexicon, morphology, and dialectal variations found in the task data.

3.3 Loss Function: Focal Loss

To address the severe class imbalance, we replaced the standard Binary Cross-Entropy (BCE) loss with **Focal Loss** (Lin et al., 2017) for the multi-label subtasks. Focal Loss modifies BCE by adding a modulating factor $(1 - p_t)^\gamma$:

$$FL(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t is the model's predicted probability for the true class. The focusing parameter γ reduces the relative loss for well-classified examples ($p_t > 0.5$), putting more focus on hard, misclassified examples. We tuned γ via cross-validation, finding $\gamma = 2.0$ optimal for the highly imbalanced English data and $\gamma = 1.5$ for the relatively more balanced Arabic data.

3.4 Two-Stage Training Strategy

A key innovation in our approach is the implementation of a two-stage training pipeline to maximize the utility of the limited training data:

1. **Stage 1 (Threshold Optimization):** We perform a stratified train/validation split (80/20). The model is trained on the 80% split, and the validation set is used to optimize the decision thresholds. Instead of a default threshold of 0.5, we calculate the optimal threshold for *each class independently* that maximises the F1-score. This is critical for rare classes where the model's output probabilities may remain consistently low (e.g., 0.3) even for positive instances.

2. **Stage 2 (Final Training):** We re-initialize the model and retrain it on the *full* training dataset (100% of the data) using the same hyperparameters. We then apply the per-class thresholds derived in Stage 1 to the final predictions on the test set. This allows the model to learn from all available data while still benefiting from calibrated decision boundaries.

4 Experimental Setup

We utilized the official SemEval-2026 Task 9 datasets for training and evaluation. All models were implemented using the Hugging Face Transformers library and PyTorch. Training was conducted on a single NVIDIA T4 GPU.

We used the AdamW optimizer with a learning rate of $2e^{-5}$ and a batch size of 16. To ensure stable convergence, we employed a linear learning rate scheduler with 10% warmup steps. The maximum sequence length was set to 128. For regularization, we applied a dropout rate of 0.1 for the hidden layers and attention probabilities. Early stopping with a patience of 2-3 epochs was used during Stage 1 to prevent overfitting.

5 Results and Discussion

Table 2 presents our results on the validation set using 5-fold cross-validation.

Task	Lang	Model	Macro F1
S1	Eng	twitter-roberta-base	**0.811**
	Ara	MARBERTv2	**0.795**
S2	Eng	DeBERTa-v3-base	**0.397**
	Ara	MARBERTv2	**0.602**
S3	Eng	twitter-roberta-base	**0.501**
	Ara	MARBERTv2	**0.567**

Table 2: Validation Set Macro-F1 Scores.

The results indicate that our Arabic models consistently outperform the English models on the fine-grained classification tasks (S2 and S3). This performance gap is likely attributable to the closer domain match between MARBERT's pre-training data (tweets) and the dataset, whereas DeBERTa is trained on more formal English text. Additionally, the extreme class imbalance in the English dataset (16:1) proved more difficult to overcome than the moderate imbalance in Arabic (2.8:1), despite the use of Focal Loss.

5.1 Ablation Studies

To quantify the impact of our design choices, we performed an ablation study on the Arabic Subtask 3 validation set (Table 3).

Configuration	Macro F1	Δ
Final System	**0.5671**	-
(A) w/o Focal Loss (BCE)	0.5457	-3.77%
(B) w/o Threshold Tuning	0.5366	-5.378%
(C) w/o Domain Pre-training	0.4774	-15.82%

Table 3: Ablation study on Arabic Subtask 3. "Final System" uses Focal Loss + Threshold Tuning + MARBERTv2.

Impact of Domain-Adaptive Pre-training: Configuration (C) shows the most dramatic drop in performance (-15.82%). Replacing the tweet-based MARBERTv2 with a generic multilingual model (mBERT) severely degrades results. This confirms that for dialectal Arabic social media analysis, coverage of informal morphology and lexicon is far more critical than multilingual capacity.

Impact of Loss Function: Comparing the Final System with Configuration (A), removing Focal Loss resulted in a nearly 4% drop in F1. While standard BCE maintained high accuracy, it did so by sacrificing recall on minority classes. Focal Loss successfully balanced the gradients, allowing the model to learn representations for rare manifestations.

Impact of Threshold Tuning: Configuration (B) highlights the value of our dynamic thresholding strategy. Using a fixed 0.5 threshold is suboptimal because model confidence varies by class frequency. Optimizing thresholds recovered roughly 5.4% in F1 score, proving it to be a cost-effective post-processing step.

6 Negative Results and Error Analysis

In the spirit of scientific transparency, we report approaches that failed to generalize in our experiments.

Model Size vs. Data Scarcity: We attempted to fine-tune **DeBERTa-v3-Large**. Despite its superior performance on GLUE benchmarks, the Large model achieved a lower Validation F1 (0.35) on English Subtask 2 compared to the Base model (0.39). We observed that the training loss converged to near-zero within 2 epochs, indicating rapid overfitting. This suggests that for small, specialized datasets ($N < 3000$), the regularization benefits of smaller models outweigh the capacity of larger ones.

Ensemble Learning: We experimented with a weighted ensemble of the 5-fold cross-validation models. This yielded only a marginal improvement of F1. Given the 5x increase in inference latency and computational cost, we deemed this trade-off inefficient for the final system submission.

7 Conclusion

Our participation in SemEval-2026 Task 9 highlights that in low-resource, imbalanced scenarios, architectural appropriateness and loss function engineering are more critical than raw model scale. By tailoring the loss function to the data distribution via Focal Loss and leveraging domain-adaptive pre-training with MARBERTv2 and Twitter-RoBERTa, we achieved competitive performance. Our comprehensive EDA revealed that the 16:1 class imbalance and multi-label complexity necessitated specialized techniques beyond

standard classification approaches. Future work will focus on investigating data augmentation techniques to further address the scarcity of minority class examples.

Acknowledgments

We thank the organizers of SemEval-2026 for providing the datasets and the baseline code structure. We also thank Dr. Shamsuddeen Muhammad and Dr. Idris Abdulmumin for their valuable guidance and feedback regarding the implementation of this project.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. 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 7088–7105, Online. Association for Computational Linguistics.

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.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

A Code Availability

To ensure reproducibility, we have made our entire system code publicly available. This includes the data preprocessing scripts, the two-stage training implementation, and the evaluation notebooks used to generate the results in this paper.

The code repository can be accessed at: `https://github.com/Mohammed-Nagi/NLP.git`