PolarizedTeam at SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization

Nestor Maria[1] Al shrafat Maroan[1] Pește Ioana[1] Daniela Gifu[2,3] Diana Trandabăț[1]

[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași,
[2]Institute of Computer Science, Romanian Academy- Iasi Branch,
[3]Academy of Romanian Scientists,

{nesmariaes, maroan241, impeste}@gmail.com

daniela.gifu@iit.academiaromana-is.ro

diana.trandabat@info.uaic.ro

Abstract

This paper presents the systems developed for SemEval-2026 Task 9, which targets the detection and categorization of multilingual, multicultural, and multi-event online polarization across 22 languages. To address the challenges posed by linguistic diversity and short, heterogeneous texts, we evaluate several Transformer-based architectures for multilingual polarization detection. Our approach models the task as a multi-label classification problem and incorporates mean pooling for sentence representation, focal loss to mitigate severe label imbalance, and label-wise attention mechanisms to capture polarization-specific linguistic cues. Experimental results show that combining robust multilingual encoders with label-aware modelling substantially improves the detection of polarized content across diverse communities and events[1].

1 Introduction

Online polarization has become a pervasive phenomenon in contemporary digital communication, manifesting through escalating antagonism between social, political, cultural, or identity-based groups. Prior research has shown that polarized discourse often precedes more severe forms of harmful online behavior, including hate speech, harassment, and misinformation, with measurable spillover effects into offline social tensions (Warmsley, 2017). Detecting polarization early is therefore essential for enabling timely interventions, content moderation strategies, and large-scale monitoring of discursive dynamics across communities and events.

The SemEval-2026 Task 9 introduces a multilingual, multicultural, and multievent benchmark designed to capture polarization across 22 languages, reflecting the increasing need for robust cross-lingual models capable of handling heterogeneous, short, and noisy online texts. Despite substantial progress in multilingual NLP, several open challenges remain. Transformer-based encoders such as mBERT and XLM-RoBERTa have demonstrated strong cross-lingual transfer capabilities, yet they struggle with extreme label imbalance, sparse contextual cues in short texts, and the need to capture label-specific linguistic signals, particularly in multi-label settings where different types of polarization may co-occur. Moreover, the diversity of languages included in the POLAR dataset, ranging from high-resource to severely low-resource languages, raises a legitimate question: *To what extent can multilingual encoders, enhanced with task-specific modelling strategies, reliably detect fine-grained polarization across heterogeneous linguistic, cultural, and event-driven contexts?*

This question guides our system design and evaluation.

Our contributions are threefold. (1) We conduct a systematic comparison of multilingual Transformer-based architectures for polarization detection across 22 languages, highlighting the strengths and limitations of mBERT, XLM-RoBERTa, and hybrid encoder–decoder configurations. (2) We introduce a modelling pipeline that integrates mean pooling for stable sentence representations, focal loss to address severe label imbalance, and label-wise attention mechanisms that capture polarization-specific linguistic cues in a multi-label setting. (3) We provide an extensive cross-lingual evaluation and error analysis, showing that label-aware modelling substantially improves performance on rare polarization types and low-resource languages.

2 Background

Research on online polarization spans multiple methodological traditions, reflecting the complex-

[1]https://github.com/FreaksMind/POLAR-SEMEVAL

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1391–1397

ity of modeling antagonistic discourse across languages, cultures, and platforms. Early computational approaches relied on supervised classification using lexical, syntactic, and surface-level features, typically employing logistic regression, support vector machines, or random forests (Warmsley, 2017).

The POLAR shared task at SemEval-2026 builds on this line of work by introducing a multilingual, multicultural, and multievent benchmark covering 22 languages. The task comprises two subtasks: Subtask 1, a binary classification task determining whether a text contains polarized content, and Subtask 2, a multi-label classification task identifying specific types of polarization (**?**).

These subtasks present significant challenges, including severe label imbalance, heterogeneous text lengths, sparse contextual cues, and substantial variation across languages.

To address these challenges, we explore multiple multilingual Transformer-based architectures. Our systems include a baseline mBERT model, two XLM-RoBERTa variants (base and large), and a hybrid architecture combining mBERT with an LSTM layer to capture sequential dependencies. Our training strategies are adapted to the characteristics of the POLAR dataset, incorporating techniques for handling class imbalance, label sparsity, pooling strategies, and threshold calibration to improve robustness across languages and polarization types.

3 Dataset and Methods

This section introduces the dataset and the methodological approach adopted in our system.

3.1 Dataset

The dataset provided for SemEval-2026 Task 9 consists of short textual snippets collected from online platforms and news or commentary forums across 22 languages: Amharic, Arabic, Bengali, Burmese, Chinese, English, German, Hausa, Hindi, Italian, Khmer, Nepali, Odia, Persian, Polish, Punjabi, Russian, Spanish, Swahili, Telugu, Turkish, and Urdu. Each language is represented by approximately 1,700 to 7,000 instances, with most languages containing around 3-4k samples.

For Subtask 1, each instance includes an ID, the text, and a binary label indicating whether the content is polarized. The overall distribution of polarized versus non-polarized instances highlights an imbalanced dataset.

For Subtask 2, each text is annotated with zero or more of five polarization categories: political, racial/ethnic, religious, gender/sexual, and other. The distribution of these categories across languages is highlighting substantial variation and extreme sparsity for several labels. To mitigate this imbalance, non-polarized instances (all-zero labels) are downsampled to 50% of the number of polarized samples. Several languages contain categories with very few or no positive examples, further motivating the use of imbalance-aware training strategies.

Overall, the dataset presents several challenges: multilingual variability, heterogeneous text lengths, severe label imbalance, and sparse positive examples for certain polarization types. These characteristics directly informed our modelling and training decisions.

3.2 Methods

Based on insights from prior work demonstrating the effectiveness of multilingual Transformer-based models, we adopted (Devlin et al., 2019) as a strong and widely used baseline across tasks, and selected XLM-RoBERTa models (Conneau et al., 2020) or their superior multilingual representations and robustness in cross-lingual settings. We describe below the architectures and training procedures adopted for Subtask 1 and Subtask 2 (see Figure 3).

3.2.1 Subtask 1

Subtask 1 is formulated as a binary classification problem, where each text is labeled as polarized or non-polarized. We evaluate three architectures: mBERT, XLM-RoBERTa-Large, and a hybrid mBERT + LSTM model.

Preprocessing and Data Handling

Texts are normalized by removing punctuation, hashtags, and non-informative symbols. Due to the lack of reliable lemmatizers for all 22 languages, no lemmatization is applied. Inputs are truncated or padded to a maximum sequence length of 256 tokens. For each language, the dataset is split into 80% training and 20% validation.

Model Architectures

1. **mBERT baseline**

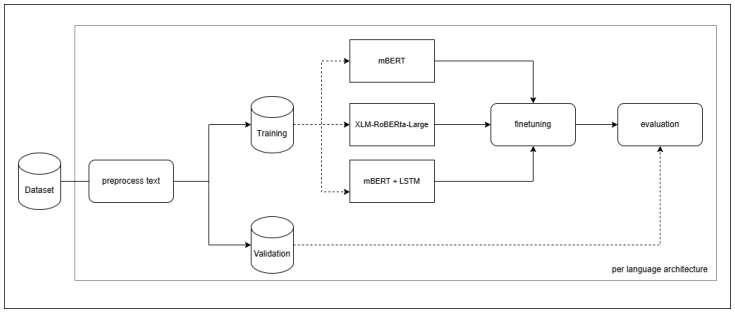

Figure 1: Architecture of the system implemented per language for Subtask 1

We fine-tune the multilingual BERT encoder (Devlin et al., 2019), which provides contextualized representations across 100+ languages. However, mBERT is known to struggle with very short texts due to limited contextual information.

2. **XLM-RoBERTa-Large**

This model extends RoBERTa to a multilingual setting and is trained on significantly larger CommonCrawl data. Its increased depth and parameter count allow it to capture richer semantic and cross-lingual patterns (Conneau et al., 2020).

2. **Hybrid mBERT + LSTM**

To complement transformer-based contextual embeddings with sequential modeling, we feed mBERT token embeddings into a unidirectional LSTM layer (Hochreiter and Schmidhuber, 1997) This hybrid architecture aims to capture sequential cues and sentiment shifts that may signal polarization. Adding an attention layer on top of the LSTM did not yield improvements and occasionally led to overfitting, as observed in our validation experiments.

3.2.2 Subtask 2

Subtask 2 requires predicting one or more of five polarization categories for each text: political, racial/ethnic, religious, gender/sexual, and other. The task is substantially more challenging than Subtask 1 due to extreme label imbalance, heterogeneous text lengths, and the presence of languages with very sparse or even missing positive examples for certain categories.

1. **Preprocessing and Label Handling**

Each instance is annotated with a five-dimensional binary vector. To mitigate the dominance of non-polarized examples (all-zero vectors), we downsample these instances to 50% of the number of polarized samples in each language. For languages where a label has zero positive instances, that label is excluded from training to avoid degenerate gradients and unstable loss behavior. All texts are tokenized using the corresponding multilingual tokenizer and truncated or padded to a maximum length of 256 tokens.

2. **Baseline Architecture**

The baseline model uses mBERT (Devlin et al., 2019) as encoder. Token embeddings are aggregated using mean pooling, which provides stable multilingual sentence representations across languages with highly variable text lengths. The pooled embedding is passed through a two-layer feed-forward classifier with ReLU activation and dropout. Training uses BCEWithLogitsLoss with class-specific positive weights computed from the training distribution to counteract label imbalance.

3. **Intermediate Architecture**

To improve sentence-level representations, we replace mBERT with XLM-RoBERTa-Base (Conneau et al., 2020), which offers stronger multilingual contextualization. We concatenate the [CLS] embedding with the mean-pooled embedding to capture both global and token-averaged information. This combined representation is projected through a feed-forward network. To further address imbalance, we adopt Focal Loss, which down-weights easy negatives and focuses learning on rare or difficult positive examples.

4. **Final Architecture: Label-Wise Attention**

To capture label-specific linguistic cues, we introduce a label-wise attention mechanism. For each polarization category, a dedicated attention head computes token-level importance scores, producing a label-specific sentence embedding. These embeddings are passed through independent linear

classifiers, enabling the model to focus on distinct lexical, semantic, and contextual patterns associated with each polarization type. This architecture consistently outperforms the baseline and intermediate models, particularly for rare labels and low-resource languages.

3.3 Experimental Setup

All experiments were implemented in PyTorch using the HuggingFace Transformers library. Models were trained independently for each language to avoid cross-lingual interference and to ensure that performance reflects language-specific distributions. We used the official training, development, and test splits provided by the SemEval-2026 Task 9 organizers.

Training Configuration
We fine-tuned all architectures using the AdamW optimizer with a learning rate of 1×10^{-5} and a linear decay schedule. Gradient clipping with a max-norm of 1.0 was applied to stabilize updates. Batch sizes below 32 produced unstable gradients, while training for more than five epochs consistently led to overfitting; therefore, we used batch sizes of 32–64 and trained for a maximum of five epochs with early stopping based on development loss.

Dropout was applied in all classifier layers to reduce overfitting, and weight decay was set to 0.01. All models used a maximum sequence length of 256 tokens, with truncation applied to longer texts and padding applied to shorter ones. No lemmatization or language-specific preprocessing was performed to maintain comparability across languages.

Intermediate System: Mean Pooling and Focal Loss
To enrich sentence-level representations, we employed XLM-RoBERTa-Base as encoder. Instead of relying solely on the [CLS] token, we computed a mean-pooled representation over all token embeddings, following evidence that mean pooling yields more stable multilingual sentence representations (Del and Fishel, 2022).

Let $H \in \mathbb{R}^{L \times d}$ be the last hidden layer of the transformer, and let $M \in \{0, 1\}^L$ denote the attention mask. The pooled representation $h \in \mathbb{R}^d$ is computed as:

$$h = \frac{\sum_{i=1}^{L} H_i * M_i}{\sum_{i=1}^{L} M_i + \epsilon} \quad (1)$$

where L is the sequence length, d is the hidden dimension, and ϵ prevents division by zero.

The final representation is obtained by concatenating the [CLS] embedding with the mean-pooled vector, combining global contextual information with aggregated token-level semantics. This representation is passed through a feed-forward network consisting of a linear layer ($768 \rightarrow 256$), ReLU activation, dropout, and a final layer producing one logit per label.

Due to extreme label imbalance, we replaced standard cross-entropy with Focal Loss, which focuses training on hard examples and down-weights easy negatives.

This intermediate system achieved an average macro-F1 of 0.495, improving over the baseline but still limited by the use of a shared representation for all labels.

Final System: Label-Wise Attention

To capture label-specific linguistic cues, we extended the architecture with a label-wise attention mechanism and independent classifiers for each polarization type. Each label receives its own attention head, producing a dedicated sentence embedding. These embeddings are passed through dropout and independent linear classification heads. Additionally, lower encoder layers were partially frozen to improve stability in low-resource languages.

This final architecture achieved an average macro-F1 of 0.507, demonstrating clear improvements over both the baseline and intermediate systems.

4 Results

This section presents the performance of the three architectures evaluated for Subtask 1 and Subtask 2 across all 22 languages. We report average accuracy and macro-F1 following the official SemEval evaluation protocol. Results are summarized in Tables 1–6, covering all languages and all model variants.

4.1 Subtask 1 Results

Performance for Subtask 1 is reported in Tables 1, 2 and 3, which present macro-F1 metric for

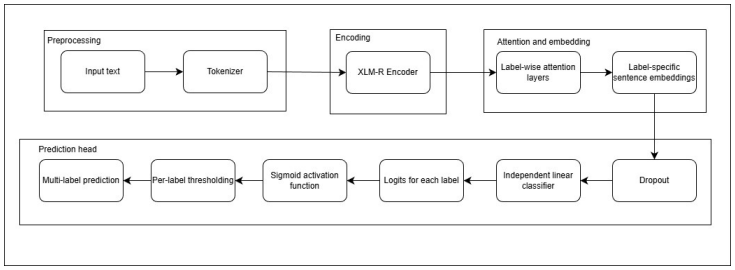

Figure 2: Final System Architecture used for Subtask 2

mBERT, XLM-RoBERTa-Large, and the hybrid mBERT+LSTM model across all languages.

Across the board, the hybrid model achieves the strongest results, outperforming both mBERT and XLM-RoBERTa-Large in most languages. This trend is visible in both Table 1 (first set of languages) and Table 2 (remaining languages), where the hybrid architecture consistently yields the highest macro-F1 values.

The mBERT baseline performs well in high-resource languages such as English, Spanish, and Persian, but struggles in languages with short or noisy texts (e.g., Khmer, Burmese), confirming its sensitivity to limited contextual cues.

XLM-RoBERTa-Large improves over mBERT in several mid-resource languages (e.g., German, Hindi, Bengali), but its performance remains inconsistent across the full multilingual spectrum, as shown in both tables.

Overall, Subtask 1 results indicate that combining contextual embeddings with sequential modeling provides the most robust cross-lingual performance.

4.2 Subtask 2 Results

Results for Subtask 2 are also presented in Tables 4, 5 and 6, covering all 22 languages. The task is substantially more challenging due to extreme label imbalance and the multi-label nature of the problem.

The baseline mBERT model achieves an average macro-F1 of 0.453, with strong performance in high-resource languages but significantly lower scores in languages with sparse positive labels.

The intermediate XLM-RoBERTa-Base model, which integrates both [CLS] and mean-pooled embeddings, improves the average macro-F1 to 0.495. This improvement is visible across both tables, particularly in Bengali, Hindi, German, and Spanish.

The final label-wise attention architecture achieves the best overall performance, with an average macro-F1 of 0.507. This model shows consistent gains across languages with highly skewed label distributions, such as Amharic, Hausa, and Telugu.

4.3 Model Comparison and Architecture Insights

Figure 2 shows the final Subtask 2 architecture, where label-wise attention assigns each label its own attention head and classifier, helping capture fine-grained cues beyond generic pooling.

Table 6 reports aggregated results across languages and models, confirming that the final architecture achieves the best macro-F1 across both subtasks. Overall, the results show that hybrid models perform best for binary polarization detection, label-wise attention yields the largest gains for multi-label classification, Focal Loss improves rare-category performance, and label-specific modeling enhances cross-lingual robustness. These trends are consistent across all 22 languages, as shown in Tables 1–6, validating the effectiveness of combining multilingual contextualization with task-specific mechanisms.

5 Discussions

Although our systems achieve consistent improvements across both subtasks, several limitations remain. The most persistent challenge is extreme label imbalance, which continues to affect rare categories despite the use of Focal Loss and label-wise attention. Performance also drops sharply in languages with very short or noisy texts, where contextual cues are insufficient for Transformer-based encoders. In addition, training models independently per language prevents cross-lingual transfer that could benefit low-resource settings. Finally, attention heads in the final architecture show higher variance across languages, suggesting sensitivity to sparse training signals. These limitations indicate that multilingual polarization detection remains dif-

Model	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
mBERT	0.442	0.669	0.679	0.362	0.745	0.714	0.742	0.672	0.567	0.475	0.53
XLM-R	0.685	0.757	0.764	0.676	0.762	0.684	0.518	0.733	0.479	0.475	0.660
Hybrid	0.727	0.770	0.840	0.671	0.762	0.837	0.811	0.784	0.648	0.585	0.858

Table 1: F1 scores per model and language (amh → mya)

Model	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
mBERT	0.509	0.417	0.632	0.605	0.627	0.695	0.774	0.655	0.630	0.654	0.689
XLM-R	0.799	0.723	0.653	0.722	0.604	0.592	0.639	0.703	0.773	0.650	0.845
Hybrid	0.880	0.752	0.740	0.729	0.741	0.696	0.799	0.872	0.713	0.760	0.897

Table 2: F1 scores per model and language (nep → zho).

Model	Average Accuracy	Average F1 Score
mBERT	0.70	0.61
XLM-RoBERTa-Large	0.71	0.67
mBERT + LSTM	0.80	0.76

Table 3: Average accuracy and F1 scores across all languages for different models.

ficult, particularly for low-resource languages, rare labels, and short-text scenarios.

6 Conclusion

This work presented a multilingual system for detecting online polarization across 22 languages, covering both binary detection and fine-grained multi-label classification. Our experiments show that combining strong multilingual encoders with task-specific strategies consistently improves over baseline Transformer models.

For Subtask 1, the hybrid mBERT+LSTM model was most effective, suggesting that sequential modelling complements contextual embeddings for short, heterogeneous texts. For Subtask 2, mean pooling, imbalance-aware optimization, and label-wise attention improved rare-category detection and achieved the highest macro-F1 across languages.

Challenges remain in low-resource languages, imbalanced label distributions, and very short or noisy texts, pointing to future work on cross-lingual transfer, data augmentation, and models suited to sparse inputs. Overall, the results show that multilingual polarization detection benefits from combining robust cross-lingual representations with mechanisms that capture label-specific linguistic cues.

Acknowledgments

This work was carried out partially within the project "Tools for Processing Online Texts Specific to Cultural and Scientific Diplomacy", funded by the Academy of Romanian Scientists.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Maksym Del and Mark Fishel. 2022. Similarity of sentence representations in multilingual lms: Resolving conflicting literature and case study of baltic languages. *Preprint*, arXiv:2109.01207.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Lo, Soda Marem and Stranisci, Marco A. and Cignarella, Alessandra Teresa and Frenda, Simona and Basile, Valerio and Bosco, Cristina and Jezek, Elisabetta and Patti, Viviana. 2025. Subjectivity in stereotypes against migrants in Italian : an experimental annotation procedure. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, volume 4112, pages 1–10. AILC.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, 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

Model	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
mBERT	0.333	0.441	0.222	0.491	0.392	0.578	0.100	0.776	0.341	0.258	0.527
XLM-R (CLS+mean pooling and Focal Loss)	0.488	0.531	0.209	0.479	0.337	0.621	0.077	0.804	0.343	0.652	0.500
XLM-R (label-wise attention)	0.511	0.518	0.287	0.503	0.343	0.592	0.089	0.801	0.361	0.661	0.551

Table 4: F1 scores per model and language (amh → mya).

Model	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
mBERT	0.742	0.149	0.385	0.379	0.431	0.606	0.470	0.437	0.436	0.758	0.704
XLM-R (CLS+mean pooling and Focal Loss)	0.713	0.481	0.345	0.439	0.419	0.583	0.462	0.468	0.449	0.751	0.734
XLM-R (label-wise attention)	0.728	0.468	0.401	0.417	0.533	0.595	0.453	0.386	0.445	0.750	0.753

Table 5: F1 scores per model and language (nep → zho).

Model	Average F1 Score
mBERT	0.45
XLM-R (CLS+mean pooling and Focal Loss)	0.49
XLM-R (label-wise attention)	0.50

Table 6: Average F1 scores across all languages.

and Multievent Online Polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. 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.

Dana Warmsley. 2017. On the detection of hate speech, hate speakers and polarized groups in online social media.

Wentao Xu, Wenlu Fan, Shiqian Lu, Tenghao Li, and Bin Wang. 2025. Polarized patterns of language toxicity and sentiment of debunking posts on social media. *Preprint*, arXiv:2501.06274.