# Joshualee2 at SemEval-2026 Task 9: Cross-Lingual Transformer-Based Polarization Detection

**Joshua Lee**[†]
[†]De Anza College

## Abstract

This paper describes our system for POLAR Subtask 1 on multilingual polarization detection. The task involves binary sequence classification over 22 languages, where the model aims to predict whether a given text exhibits polarized discourse. To deal with the multilingual and resource-imbalanced nature of the dataset, we fine-tune XLM-R, a pre-trained multilingual transformer encoder, using a language-aware sampling strategy that combines all available training data into a unified multilingual corpus. Our system achieves an overall macro-F1 of 0.781 and an average accuracy of 0.823 on the official test set. Results show strong performance in low-resource languages, though some discrepancies indicate remaining class imbalance. All code used for training and evaluation is publicly available.[1]

## 1 Introduction

The increase of polarized, non-constructive discourse on social media has raised growing concerns on its societal impact, including amplifying political division, misinformation, and online hostility. Detecting polarized content is therefore important for content moderation and is a task targeted by natural language processing researchers. POLAR at SemEval 2026 (Naseem et al., 2026b,a) focuses on multilingual polarization detection, supporting the need for systems capable of handling cross-cultural and linguistic diversity in political viewpoints and discourse. In Subtask 1, participants are asked to build systems for binary classification, to determine whether a given social media post expresses polarized opinion or not.

POLAR provided a 22-language corpus and CodaBench evaluation tools for fair cross-lingual comparison. The CodaBench evaluation tool took model outputs that were filled into a language specific CSV file, and evaluated their performance under several metrics, including accuracy, precision, and recall. These resources allowed training and evaluation of multilingual models under controlled conditions, which allowed a fair comparison across languages and approaches.

This work approaches the task using a multilingual transformer-based architecture built upon XLM-RoBERTa-large (Conneau et al., 2020). Figure 1 presents an overview of our multilingual training and classification pipeline. Our system uses cross-lingual transfer learning to learn representations across all 22 languages while accounting for dataset imbalance in training data through language sampling strategies. Specifically, to train a single multilingual model, we combined all per-language training sets into one dataset and used a balanced training procedure that reweights batches using a weighted random sampler to create equal language representation during training. We additionally use early stopping, gradient accumulation, and macro-F1 optimization to improve generalization performance across languages.

During participation, we observed that class imbalance and language distribution differences affect performance, motivating our language balancing strategy. Additionally, model performance varied notably between high-resource and lower-resource languages, highlighting the importance of cross-lingual transfer and balanced sampling when training multilingual models.

Our approach achieved an average macro-F1 score of 0.781 and an average accuracy of 0.823 across 22 languages in the official evaluation. Performance was strong in several languages such as Nepali, Chinese, Hindi, and Persian, while lower scores were observed in Khmer and Italian in relation to macro-F1 scores. These results demonstrate that large multilingual pretrained models can effectively generalize across diverse linguistic settings with appropriate training strategies, although there are still differences in performance specific to each language.
[2]

## 2 Background

Polarized discourse on social media has been linked to increased political division, misinformation, and online hostility, motivating research in polarization detection. POLAR at SemEval-2026 (Naseem et al., 2026b,a) focuses on this challenge as a multilingual classification problem, where approaches must determine whether a social media post contains a polarized opinion across different cross-cultural and linguistic contexts. Subtask

[1] https://github.com/joshualee2006164/polar-semeval-2026.git

[2] XLM-Roberta-Large: https://huggingface.co/FacebookAI/xlm-roberta-large



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 2760–2764
July 3-4, 2026 ©2026 Association for Computational Linguistics

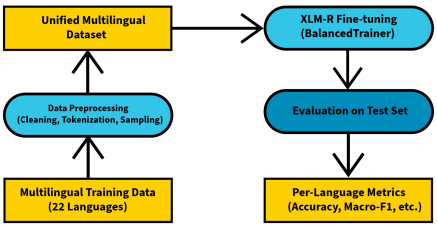


Figure 1: Overview of the proposed multilingual polarization detection system.

1 of POLAR specifically targets binary polarization detection, with labels representing either polarized or not polarized content.

Early approaches to polarization detection relied on a sentence's lexical features, which failed to capture contextual meaning and cross-lingual variation. More recent advances in representation learning have substantially improved natural language understanding across similar tasks (Mikolov et al., 2013; Peters et al., 2018; Devlin et al., 2019). Multilingual pretrained models such as mBERT and XLM demonstrated that shared semantic representations can emerge across languages without specific cross-lingual supervision, enabling transfer learning in low-resource settings (Lample and Conneau, 2019).

With these developments, large-scale multilingual models like XLM-RoBERTa achieve strong performance across diverse languages by combining a large training corpora with pretraining (Conneau et al., 2020). Using multilingual contextual representations allows models to learn shared semantic features while capturing nuances that are specific to languages.

Our approach builds on these foundations by combining XLM-RoBERTa, a multilingual transformer, with a language-aware balanced training strategy, designed to mitigate class imbalance and improve performance on low-resource languages.

## 3 System Overview

Our system detects multilingual polarization by using a transformer-based architecture built on XLM-RoBERTa-large (Conneau et al., 2020). The pipeline involves multilingual data aggregation, tokenization, balanced training with language-aware sampling, and fine-tuning with early stopping to optimize for macro-F1 performance.

### 3.1 Language Balanced Training

A key challenge in the task is imbalances across languages and label distributions. To solve this issue, we used a language-aware sampling strategy that ensures equal representation of languages during training.

Specifically, we compute sampling weights inversely proportional to the number of training instances per language:

$$w_l = \frac{1}{N_l}$$

where $N_l$ is the number of samples for a certain language $l$. These weights are used with a weighted random sampler that balances batches across languages.

### 3.2 Training Procedure

We fine-tune XLM-RoBERTa-large using the Hugging Face Transformers framework with the following hyper parameters:

- Learning rate: $1 \times 10^{-5}$
- Batch size: 16 with gradient accumulation
- Maximum epochs: 8
- Warmup ratio: 0.1
- Weight decay: 0.01
- Gradient clipping: 1.0

We optimized models for the best macro-F1 on the development set and CodaBench Polarization Detection Evaluation System. Early stopping with a patience of three epochs was applied to avoid overfitting. Our optimizer was AdamW.

### 3.3 Models

We experimented with different training configurations, including different batch sizes, learning rates, and epochs. However, the primary comparison was with random sampling and the language-balanced sampling approach. The balanced strategy consistently improved macro-F1 performance and was therefore used for the final submission.

## 4 Experimental Setup

### 4.1 Data Splits and Usage

We use the multilingual dataset provided by the POLAR shared task at SemEval-2026 (Naseem et al., 2026b), which consists of social media posts annotated for binary polarization across 22 languages.

Each split is distributed as language-specific CSV files. During training, we load all language files within a split and concatenated to create a single, unified multilingual dataset, saving the language identifiers as metadata. The development set was used only for validation, model selection, and early stopping. The best checkpoint selected on the development set (based on macro-F1) was used for the final submission.

### 4.2 Preprocessing

Minimal preprocessing was applied to our unified multilingual dataset and each individual language. The preprocessing consisted of:

**Language tagging:** Each example is assigned a language code based on their filename. This metadata is used for balanced sampling and per-language evaluation.

**Tokenization:** Text is tokenized using the tokenizer associated with XLM-RoBERTa (Conneau et al., 2020). We apply a maximum sequence length of 256 tokens with truncation for longer inputs and dynamic padding using a data collator.

**Data Processing:** The provided POLAR dataset has separate CSV files for each language. We merge all languages into a single training dataset, using the language identifiers as metadata.

### 4.3 Evaluation Metrics

Our evaluation metrics consisted of a Macro-averaged F1 score (primary metric), Accuracy, and Per-language F1 and accuracy for diagnostic analysis

Macro-F1 is computed across the two classes to provide balanced evaluation regardless of class distribution.

### 4.4 Test Prediction Generation

For the test phase, predictions are generated separately for each language file to match the submission format required by the evaluation platform.

## 5 Results

### 5.1 Official Evaluation Results

Our final submission was evaluated using the official POLAR SemEval-2026 test set which covers 22 languages using macro-averaged F1 as the primary metric. The detailed per-language results are shown in Table 1. Our system scored an overall macro-F1 score of **0.781** and an average accuracy of **0.823**.

### 5.2 Per-Language Performance

Performance varies across languages, as shown in Table 1, reflecting differences in dataset size and linguistic characteristics per language. High-resource languages such as Chinese (0.882 macro-F1) and Burmese (0.876) show strong performance. Low-resource languages, such as Nepali (0.908) and Urdu (0.818), also achieve strong macro-F1 scores, suggesting that our language-aware sampling strategy helps with data scarcity. In contrast, languages such as Italian (0.647), German (0.704), and Khmer (0.575) exhibit lower macro-F1 scores.

Notably, some languages with very high accuracy (e.g., Khmer and Hausa) show substantially lower macro-F1 scores, indicating class imbalance effects where the model favors the majority class. This highlights the importance of macro-F1 as an evaluation metric for balanced performance across classes.

### 5.3 Key Findings

Several important trends emerge from the results:

**Benefits for low-resource languages:** The language-aware sampling strategy appears to improve performance for several low-resource languages, particularly Nepali and Urdu, which achieve strong macro-F1 scores.

**Impact of class imbalance:** Languages with skewed label distributions exhibit larger gaps between accuracy and macro-F1, suggesting that class imbalance remains a challenge despite balanced sampling.

### 5.4 Discussion

While we do not have access to additional computational resources to perform post-hoc evaluations, we can observe for languages such as Hausa (HA) and Khmer (KH), there is a large gap between the precision and recall. This indicates that the model may over-predict one class over another. The large gap between the F1 score for binary and macro F1 implies that class imbalance is affecting the model.

Overall, errors are likely due to a combination of low-resource conditions, language-specific nuances, and subjectivity of polarization annotations. Future work could include manual inspection of false positives and false negatives, error categorization by linguistic features, and targeted model adaptation for low-resource languages to improve robustness.

### 5.5 Analysis Beyond Overall Scores

The discrepancy between accuracy and macro-F1 for certain languages indicates that future improvements should focus on better class imbalance handling, language-specific adjustments, and improved modeling of polarization cues.

Despite these challenges, the system demonstrates strong generalization across languages, validating the effectiveness of large multilingual transformer models for polarization detection without extensive task-specific engineering.

## 6 Ethical Considerations

Polarization detection systems operate in sensitive domains, particularly when applied to political or ideological discourse. While the goal of this task is to understand polarized language, models could be misused for large-scale surveillance, censorship, or suppression of political expression, leading to unfair moderation of content.

Finally, while the task uses publicly available social media data and no additional personal information is collected, the information can be sensitive, calling for ethical data handling. To reduce these risks, polarization detection systems should be used only as assistive tools and should be tested on different demographics and linguistic groups.

## 7 Conclusion

This paper presented our system for Subtask 1 (binary polarization detection) of the POLAR shared task at

| Language | Accuracy | Precision | Recall | F1 Binary | F1 Macro | F1 Micro |
|---|---|---|---|---|---|---|
| Amharic (AM) | 0.8281 | 0.8588 | 0.9178 | 0.8873 | 0.7625 | 0.8281 |
| Arabic (AR) | 0.8212 | 0.8030 | 0.7959 | 0.7994 | 0.8190 | 0.8212 |
| Bengali (BE) | 0.8314 | 0.7941 | 0.8104 | 0.8022 | 0.8277 | 0.8314 |
| German (DE) | 0.7046 | 0.6948 | 0.6837 | 0.6892 | 0.7039 | 0.7046 |
| English (EN) | 0.8003 | 0.7306 | 0.7223 | 0.7264 | 0.7846 | 0.8003 |
| Persian (FA) | 0.8524 | 0.8839 | 0.9217 | 0.9024 | 0.7998 | 0.8524 |
| Hausa (HA) | 0.9148 | 0.6733 | 0.3886 | 0.4928 | 0.7231 | 0.9148 |
| Hindi (HI) | 0.9005 | 0.9370 | 0.9468 | 0.9418 | 0.7987 | 0.9005 |
| Italian (IT) | 0.6671 | 0.7455 | 0.4505 | 0.5616 | 0.6467 | 0.6671 |
| Khmer (KH) | 0.9143 | 0.9168 | 0.9959 | 0.9548 | 0.5749 | 0.9143 |
| Burmese (MY) | 0.8793 | 0.8858 | 0.9060 | 0.8958 | 0.8762 | 0.8793 |
| Nepali (NE) | 0.9081 | 0.9017 | 0.9157 | 0.9087 | 0.9081 | 0.9081 |
| Odia (OR) | 0.8302 | 0.7103 | 0.6799 | 0.6948 | 0.7886 | 0.8302 |
| Punjabi (PA) | 0.7738 | 0.7442 | 0.8142 | 0.7776 | 0.7737 | 0.7738 |
| Polish (PO) | 0.8106 | 0.7947 | 0.7384 | 0.7655 | 0.8033 | 0.8106 |
| Russian (RU) | 0.8203 | 0.6984 | 0.7000 | 0.6992 | 0.7855 | 0.8203 |
| Spanish (SP) | 0.7661 | 0.7416 | 0.8082 | 0.7734 | 0.7659 | 0.7661 |
| Swahili (SW) | 0.7474 | 0.7696 | 0.7087 | 0.7379 | 0.7470 | 0.7474 |
| Telugu (TE) | 0.8555 | 0.8350 | 0.8986 | 0.8656 | 0.8547 | 0.8555 |
| Turkish (TU) | 0.7731 | 0.7978 | 0.7557 | 0.7762 | 0.7731 | 0.7731 |
| Urdu (UR) | 0.8182 | 0.8495 | 0.8968 | 0.8725 | 0.7779 | 0.8182 |
| Chinese (ZH) | 0.8822 | 0.9196 | 0.8415 | 0.8788 | 0.8821 | 0.8822 |
| **Average** | 0.8230 | 0.8180 | 0.7270 | 0.7930 | 0.7810 | 0.8370 |

Table 1: Binary classification results for Subtask 1: Polarization Detection across 22 languages for test set performance. The primary metric for evaluation is Macro F1.

SemEval-2026. Our proposed system utilized XLM-RoBERTa, a multilingual transformer model, in addition to a balanced language-aware training strategy to overcome the issue of class imbalance, which is a major problem in machine learning tasks, especially in the case of binary classification tasks. The system achieved consistent performance across 22 languages, demonstrating the effectiveness of multilingual contextual representations for polarization detection in diverse linguistic settings. Our results highlight that balancing strategies and multilingual pretraining can benefit low-resource languages, although performance variability across languages remains a challenge. Error analysis suggests that high and low resource languages significantly influence predictions. In future work, we plan to explore data augmentation techniques, improved handling of class imbalance, and parameter optimizations for multilingual methods to further improve cross-lingual generalization.

## 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. 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 NAACL*, pages 4171–4186.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 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, Xintong Wang, Surendrabikram

Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*, pages 2227–2237.

# A Code and Resources

Training and evaluation scripts for our multilingual polarization detection system are publicly available at: `https://github.com/joshualee2006164/polar-semeval-2026.git`.

The repository includes:

- **Training script:** `subtask_1_XLM-Roberta_2_git.py` for fine-tuning XLM-RoBERTa
- **Requirements:** 'requirements.txt' with Python dependencies
- **README:** Instructions for setting up environment and running experiments

The dataset is not included in the repository; please download the official POLAR shared task data from SemEval-2026.

## A.1 External Libraries and Tools

No additional datasets were used in our approach. Our system relies on publicly available resources including PyTorch for model training, the Hugging Face Transformers and Datasets libraries for model loading and data processing, scikit-learn for evaluation metrics, and Comet ML for experiment tracking and logging.