AI4PC-Howard University at SemEval-2026 Task 9: Evaluating Teacher–Student Weak Supervision and Direct LLM Prompting for Multilingual Political Polarization Detection

Surangana Aryal and Saurav K. Aryal
Howard University
Washington, DC, USA
suranagana.aryal@bison.howard.edu
saurav.aryal@howard.edu

Abstract

We describe the AI4PC–Howard University submission to SemEval-2026 Task 9, Subtask 1 on multilingual political polarization detection across 22 languages. We investigated two approaches: (1) a weakly supervised teacher–student framework in which a large language model (LLM) generated pseudo-labels to train an XLM-RoBERTa-base classifier, and (2) a context-engineered prompt-based approach using Meta-Llama-3.1-8B-Instruct. The teacher–student approach exhibited instability under distribution shift and collapsed toward majority predictions at test time. Consequently, our final submission used direct inference with Meta-Llama-3.1-8B-Instruct. While this approach produced competitive macro-F1 across evaluated languages, results reveal strong positive-class bias and substantial precision–recall imbalance. Our findings highlight limitations of weak supervision for subjective political tasks and underscore trade-offs between scalability, bias, and computational cost in LLM-only multilingual systems.

1 Introduction

Political polarization in online discourse presents challenges for democratic deliberation and cross-cultural communication (**?**Ngueajio et al., 2025). Automatically detecting polarized rhetoric enables large-scale empirical analysis of political communication across languages and regions (Waseem and Hovy, 2016). However, polarization is inherently subjective and culturally situated, making multilingual generalization difficult (**?**).

SemEval-2026 Task 9, Subtask 1 frames polarization detection as binary classification across 22 typologically diverse languages (Naseem et al., 2026a; Aryal et al., 2023a). This setting introduces challenges including domain shift, class imbalance, and cross-lingual variation in rhetorical norms. The dataset (Naseem et al., 2026b) and task formulation are described in detail by the shared task organizers (Naseem et al., 2026a), which we follow in our experimental setup.

We explore two modeling strategies:

- Teacher–student learning has been widely explored in multilingual and sentiment classification settings, but prior work shows that performance can degrade under distribution shift when pseudo-label quality is inconsistent (Aryal et al., 2023b).

- Prior work using transformer-based multilingual systems suggests that direct zero-shot or prompt-based approaches can be competitive, but often require careful calibration to avoid bias amplification across languages (Prioleau and Aryal, 2023a; Ngueajio et al., 2025).

Our experiments demonstrate that weak supervision using pseudo-labels is unstable under distribution shift, whereas direct LLM inference provides more consistent (though biased and computationally expensive) performance across evaluated languages.

2 Task Description

The task involves binary classification of short political text snippets as polarized (1) or non-polarized (0). Evaluation uses macro-averaged F1.

Although the benchmark includes 22 languages, computational constraints limited our experiments to 14 languages spanning multiple scripts and typological families. Official macro-F1 scores were computed by the organizers post-submission. Since our submission included predictions for 14 languages only, our reported results correspond to this subset and are not directly comparable to full leaderboard rankings.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2506–2511
July 3-4, 2026 ©2026 Association for Computational Linguistics

3 System Overview

We explored two architectures:

3.1 Teacher–Student Weak Supervision

Teacher Model: Meta-Llama-3.1-8B-Instruct (8B parameters), GGUF checkpoint `Meta-Llama-3.1-8B-Instruct-Q6_K.gguf`, released under the Llama 3 Community License. The teacher generated binary pseudo-labels and English rationales.

Student Model: XLM-RoBERTa-base (270M parameters), fine-tuned using cross-entropy loss. Recent work on multilingual sentiment and code-switching modeling shows that transformer-based encoders remain strong baselines across low-resource and multilingual settings (Aryal et al., 2023a,b; Prioleau and Aryal, 2023b). Prior evaluations of transformer architectures for language identification further motivate our choice of XLM-RoBERTa as a multilingual backbone (Prioleau and Aryal, 2023a).

Gold labels were replaced with teacher-generated pseudo-labels during training. Teacher predictions and rationales were available during training but not at test time. We did not fine-tune XLM-RoBERTa-base directly on gold labels. This decision was primarily driven by computational and time constraints during the shared task timeline. As a result, we do not include a supervised encoder baseline, which limits direct comparison with standard fine-tuning approaches.

It is important to note that the teacher–student pipeline was explored independently and was not used for final submission. The final system relies solely on direct LLM inference, and no pseudo-labels were reused across approaches.

3.2 Context Engineering and Prompt-Based LLM Inference (Final Submission)

Our final system uses Meta-Llama-3.1-8B-Instruct for direct binary classification without additional fine-tuning. Inference was performed using deterministic decoding (temperature 0). A fixed decision rule was used across all languages, and no per-language threshold tuning or calibration was performed. Rather than relying on zero-shot raw text classification alone, we explicitly engineered structured input context combining raw text, normalized text, and extracted stylistic features to guide model reasoning.

The exact prompt template is provided in Appendix A.

4 Stylistic Feature Extraction

In addition to raw text, we extracted language-agnostic stylistic features:

- **All-caps words:** number of tokens written entirely in uppercase.
- **Caps ratio:** proportion of uppercase characters to total characters.
- **Exclamation count:** frequency of "!" characters.
- **Question mark count:** frequency of "?" characters.
- **Multi-exclamation / multi-question:** presence of repeated punctuation sequences (e.g., "!!", "??").
- **Mixed punctuation:** presence of alternating punctuation patterns such as "!?" or "?!".
- **Elongated punctuation:** repeated punctuation beyond two characters (e.g., "!!!").
- **Repeated token bigrams:** count of consecutive repeated tokens.
- **Repeated character runs:** number of character repetitions (e.g., "soooo").
- **Maximum character run length:** longest repeated-character sequence.
- **Digit ratio:** proportion of numeric characters.
- **Punctuation ratio:** proportion of punctuation characters.
- **Unique token ratio:** vocabulary diversity.

Similar stylistic and feature-based approaches have been used in prior multilingual NLP studies, particularly in sentiment and code-switching analysis tasks, where surface-level cues provide additional signal for classification (Aryal et al., 2023a,b). Case-based features apply primarily to scripts with uppercase distinctions. No per-language normalization was performed.

We did not conduct controlled ablation experiments; thus, stylistic features are hypothesized to support reasoning but their isolated contribution is not empirically verified.

5 Why Weak Supervision Failed

Prior work in multilingual classification and biomedical NLP also highlights that weak supervision can amplify teacher bias when applied to heterogeneous or cross-domain data (Aryal et al., 2022; Prioleau and Aryal, 2023b). Contributing factors include:

- **Rationale leakage:** Student models may learn patterns from teacher rationales unavailable at inference.
- **Overfitting to teacher artifacts:** Stylized teacher outputs may introduce spurious correlations.
- **Distribution mismatch:** Teacher labels were generated on a balanced subset while test distributions differed.
- **Amplification of teacher bias:** Small systematic teacher biases may compound during fine-tuning.

This instability illustrates risks of pseudo-labeling in subjective political classification tasks. During development, we observed that the student model increasingly favored majority-class predictions on validation data, indicating instability under distribution shift even before evaluation on the test set.

6 Results and Analysis

Table 1 reports per-language performance for the 14 evaluated languages.

Positive-class bias patterns have been previously observed in multilingual sentiment and toxicity detection systems, particularly when trained on imbalanced or pseudo-labeled data (Ngueajio et al., 2025)., with recall approaching 1.0 but substantially lower precision. This imbalance results in moderate or low macro-F1 despite high recall.

For example:

- English and Russian show near-perfect recall but limited precision.
- Telugu demonstrates more balanced precision and recall, but lower overall recall.

These patterns indicate over-prediction of polarization in some languages rather than uniformly balanced classification. We also observe variation across languages that can be partially explained by dataset characteristics. For instance, Punjabi exhibits lower accuracy but higher macro-F1 compared to Hindi, likely due to more balanced class distributions. In contrast, Urdu achieves higher accuracy, which may reflect skew toward the positive class. Chinese performance remains moderate despite larger available data, suggesting potential domain mismatch or differences in linguistic expression of polarization.

We therefore avoid claiming robust multilingual generalization; performance varies substantially across languages.

7 Baseline Comparisons

We compare our system against two simple baselines computed using the true test-set label distributions.

Majority baseline: predicts the most frequent class for each language. Under binary classification, this results in zero recall for the minority class, yielding macro-F1 equal to half the F1 score of the majority class.

Random baseline: predicts each class with equal probability (0.5). Expected macro-F1 depends on class balance and approaches 0.50 only under perfectly balanced distributions.

Table 2 reports per-language baseline results.

Our LLM-based system outperforms the majority baseline in most languages, though performance remains close to baseline in highly imbalanced settings. Notably, performance on Russian falls below the majority baseline, reflecting strong positive-class bias and highlighting instability under distribution shift.

8 Related Work

Multilingual political text classification has traditionally relied on supervised fine-tuning of transformer encoders such as XLM-RoBERTa (Conneau et al., 2020). Joint multilingual fine-tuning leverages shared cross-lingual representations to improve transfer across languages, while adapter-based methods introduce lightweight language-specific modules (Houlsby et al., 2019)

Meta-learning approaches have been explored for cross-lingual domain shift (**?**) training models to adapt quickly to new languages or distributions with minimal supervision. These methods explicitly optimize for transfer robustness rather than in-language accuracy alone.

Language	Acc	Prec	Rec	F1 (Bin)	F1 (Macro)
Arabic	0.503	0.471	0.987	0.638	0.423
German	0.566	0.525	0.974	0.682	0.499
English	0.469	0.410	1.000	0.581	0.427
Persian	0.707	0.735	0.923	0.818	0.534
Hindi	0.810	0.841	0.946	0.891	0.584
Nepali	0.600	0.571	0.863	0.688	0.566
Punjabi	0.610	0.563	0.766	0.649	0.605
Russian	0.383	0.336	1.000	0.502	0.346
Spanish	0.539	0.525	0.988	0.686	0.411
Swahili	0.559	0.533	0.937	0.679	0.486
Telugu	0.525	0.532	0.424	0.472	0.521
Turkish	0.583	0.539	1.000	0.700	0.507
Urdu	0.735	0.755	0.919	0.829	0.617
Chinese	0.594	0.559	0.963	0.707	0.522

Table 1: Per-language performance of the LLM-only system.

Language	Majority Label	Majority %	Majority Macro-F1	Random Macro-F1
Arabic	0	55.3	0.356	0.495
German	0	52.4	0.344	0.499
English	0	62.6	0.384	0.484
Persian	1	74.1	0.426	0.456
Hindi	1	85.5	0.461	0.404
Nepali	1	50.3	0.334	0.500
Punjabi	0	50.6	0.336	0.500
Russian	0	69.4	0.409	0.469
Spanish	1	50.2	0.334	0.500
Swahili	1	50.1	0.334	0.500
Telugu	1	53.9	0.351	0.497
Turkish	0	51.2	0.338	0.500
Urdu	1	69.5	0.409	0.469
Chinese	0	50.4	0.335	0.500

Table 2: Majority and random baselines computed from test-set label distributions.

Translation-based pipelines represent another strategy (Artetxe et al., 2017): low-resource languages are translated into English and processed using high-resource classifiers. While effective in some settings, such pipelines introduce translation noise and increase computational overhead.

In contrast, zero-shot LLM inference bypasses supervised fine-tuning entirely (Brown et al., 2020), relying on large-scale pretraining and instruction tuning. This approach reduces engineering complexity but may exhibit bias, instability, and high inference cost. Our results illustrate both the promise and limitations of this paradigm.

9 Compute and Scalability

Each inference instance required approximately 600–900 tokens including prompt and model output. We used Meta-Llama-3.1-8B-Instruct with a 2048-token context window and deterministic decoding (temperature 0). Inference with an 8B-parameter LLM is substantially more computationally expensive than fine-tuning or running XLM-RoBERTa-base, particularly when applied independently per instance. Scaling this approach to all 22 languages increases runtime proportionally, highlighting a trade-off between engineering simplicity and computational cost.

10 Limitations

This work has several limitations. First, experiments were conducted on 14 of the 22 task languages due to computational constraints, limiting claims about full multilingual coverage. Second, we did not train a supervised XLM-R model directly on gold labels, restricting comparison against strong encoder baselines. Third, class imbalance and the absence of per-language calibration likely contributed to observed precision–recall imbalance and positive-class bias. Fourth, LLM-based inference incurs substantially higher computational cost than encoder-based models. Finally, we did not conduct ablation studies isolating the contribution of stylistic features, so their individual impact remains unverified. Ad-

ditionally, the absence of per-language calibration likely contributed to the observed precision–recall imbalance.

11 Conclusion

We presented the AI4PC–Howard University system for multilingual political polarization detection, building on prior work in multilingual sentiment modeling and transformer-based classification in low-resource settings (Aryal et al., 2023a; Prioleau and Aryal, 2023a). Weak supervision via teacher–student training proved unstable under distribution shift, highlighting risks of pseudo-labeling for subjective political tasks. Context-engineered prompt-based LLM inference offered competitive macro-F1 across evaluated languages but exhibited strong positive-class bias and substantial computational cost. Future work should explore calibration strategies such as threshold tuning, few-shot prompting with balanced examples, and probability calibration methods to mitigate positive-class bias. Future work should investigate hybrid supervised–LLM methods and improved calibration strategies for multilingual political classification.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023a. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.

Saurav K Aryal, Howard Prioleau, Surakshya Aryal, and Gloria Washington. 2023b. Baselining performance for multilingual codeswitching sentiment classification. *Journal of Computing Sciences in Colleges*, 39(3):337–346.

Saurav K Aryal, Howard Prioleau, and Gloria Washington. 2022. Sentiment classification of code-switched text using pre-trained multilingual embeddings and segmentation. *arXiv preprint arXiv:2210.16461*.

Tom Brown and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Neil Houlsby and 1 others. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

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 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.

Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys*, 57(7):1–37.

Howard Prioleau and Saurav K Aryal. 2023a. Benchmarking current state-of-the-art transformer models on token level language identification and language pair identification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 193–199. IEEE.

Howard Prioleau and Saurav Keshari Aryal. 2023b. Feature importance analysis for mini mental status score prediction in alzheimer's disease.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL*.

A Final Prompt Template

```
You are annotating texts for political
    polarization (binary label 0/1).

You are given:
1) RAW_TEXT: the original text (may
    contain casing, punctuation, emojis)
2) CLEAN_TEXT: a normalized version of
    the same text
3) STYLE_FEATURES_JSON: numeric
    stylistic and structural features
    extracted from the text
The text may be in ANY language (e.g.,
    English, Hindi, Nepali, Swahili).

How to decide:
- Polarized texts often show hostility,
    blame, us-vs-them framing,
    absolutist or extreme claims, and
    inflammatory tone
```

```
- Use STYLE_FEATURES_JSON as supporting
    signals, not as strict rules
- DO NOT rely on English-specific
    keywords
- Return reasoning IN ENGLISH

Return EXACTLY one JSON object and
    NOTHING else:
{"prediction": 0 or 1, "reasoning":
    "one short English sentence"}

RAW_TEXT:
{RAW_TEXT}

CLEAN_TEXT:
{CLEAN_TEXT}

STYLE_FEATURES_JSON:
{FEATS}

JSON:
```