DUTH at SemEval-2026 Task 9: Joint Multilingual Fine-Tuning for Online Polarization Detection

Georgios Arampatzis **Avi Arampatzis**

Department of Electrical & Computer Engineering,
Democritus University of Thrace
{geoaramp,avi}@ee.duth.gr

Abstract

Online polarization on social media presents substantial challenges for public discourse, content moderation, and large-scale social analytics across diverse linguistic and cultural contexts. A recent multilingual benchmark enables systematic evaluation of polarization detection across 22 languages and multiple sociopolitical events, providing a unified setting for studying socially grounded NLP under multilingual conditions.

We present **DUTH**, a unified multilingual system for binary polarization detection based on joint fine-tuning of XLM-RoBERTa on the 22 languages of SemEval-2026 Task 9 Subtask 1. Our system uses a single shared encoder with a linear classification head and is trained jointly on the multilingual training set using mixed-precision optimization. On the official evaluation, the system achieved an average Accuracy of 0.822 and an average Macro-F1 of 0.780 across 22 languages. The results show that a simple jointly fine-tuned multilingual transformer provides a competitive and scalable baseline for online polarization detection, while still facing difficulties in implicit, sarcastic, and culturally grounded cases.

1 Introduction

Online polarization—the growing division and antagonism between social, political, and identity groups—has become a defining challenge of contemporary digital discourse. Social media platforms amplify polarized narratives, reinforce ideological echo chambers, and facilitate the spread of inflammatory or misleading content, often intensifying intergroup hostility (Sunstein, 2009; Garimella et al., 2018). In its extreme manifestations, polarization undermines constructive dialogue and social cohesion, frequently preceding harmful online behaviors such as harassment and hate speech. As a result, the automatic detection of polarized content has emerged as a critical problem in natural language processing (NLP).

Earlier computational studies have examined polarization and related phenomena through stance detection, sentiment analysis, and network-based modeling (Conover et al., 2011). Despite substantial progress, most existing resources and systems remain concentrated on high-resource languages and narrow sociopolitical settings, limiting their applicability to multilingual and multicultural online discourse. This is particularly problematic given that polarization is often expressed through culturally grounded rhetoric, local events, and implicit cues that vary across linguistic communities.

Recent advances in multilingual pretrained language models have made it possible to address socially grounded classification tasks across typologically diverse languages using a single shared encoder (Pires et al., 2019; Conneau et al., 2020). In parallel, prior work on stance detection, affective analysis, and harmful-content detection has shown that transformer-based models can capture subtle and context-dependent cues, although performance often degrades in low-resource and culturally specific settings. Our broader line of work has similarly explored multilingual and socially grounded NLP through tasks such as intimacy analysis, sexism detection, news credibility support, justification retrieval, multilingual passage retrieval, topic-oriented retrieval, and hallucination-aware evaluation (Arampatzis et al., 2023, 2025c,a,b; Arampatzis and Arampatzis, 2025; Arampatzis et al., 2025d).

To facilitate systematic progress in polarization research across languages and cultures, the POLAR benchmark introduces a large-scale multilingual, multicultural, and multi-event dataset spanning 22 languages, with annotations covering binary polarization detection as well as fine-grained categorization of polarization types and rhetorical manifestations (Naseem et al., 2026b). This setting

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 599–604

enables rigorous evaluation of multilingual robustness and highlights the challenges posed by implicit polarization cues, sarcasm, and culturally specific discourse.

In this paper, we focus on the binary polarization detection setting and present a unified multilingual system based on joint fine-tuning of XLM-RoBERTa on all available training languages. Our goal is not to propose a novel architecture, but to assess how far a standard multilingual fine-tuning setup can go in this task when applied consistently across a large and diverse language set. We therefore emphasize reproducibility, per-language analysis, and qualitative error inspection in order to better understand the strengths and limitations of this baseline-style approach.

On the official evaluation, our system achieved an average Accuracy of 0.822 and an average Macro-F1 of 0.780 across 22 languages. The results indicate that jointly fine-tuned multilingual transformers can provide competitive performance across diverse linguistic settings, while still facing challenges in cases involving implicit polarization, sarcasm, and culturally grounded references.

The remainder of this paper is organized as follows: Section 2 reviews related work on polarization detection and multilingual NLP; Section 3 presents the system architecture and modeling approach; Section 4 describes the experimental setup and evaluation protocol; Section 5 reports the experimental results and analysis; and Section 6 concludes the paper.

2 Related Work

Polarization in Online Discourse. Early research on online polarization focused on network structures, revealing ideological segregation and the formation of echo chambers in social media interactions (Conover et al., 2011; Garimella et al., 2018). These findings established polarization detection as a supervised NLP problem grounded in both social structure and textual signals.

Neural Models for Polarization and Stance Detection. Deep learning approaches have significantly advanced the modeling of ideological and stance-related phenomena. Recurrent architectures and, more recently, transformer-based models have been employed to capture contextual and implicit polarization cues (Mohammad et al., 2016). Pretrained language models further improved robustness across topics, though most existing systems remain centered on English and other high-resource languages.

Split	Instances	Lang.	Task
Training Set	∼73,000	22	Binary
Development Set	Per language	22	Binary
Test Set	Per language	22	Binary

Table 1: Summary of the multilingual dataset used for polarization detection (Subtask 1).

Multilingual Representation Learning. Multilingual pretrained transformers have enabled scalable text classification across languages by learning shared semantic representations. Models such as XLM-RoBERTa exhibit strong performance in diverse NLP tasks, particularly when jointly fine-tuned across multiple languages (Conneau et al., 2020). This paradigm has proven effective for sentiment analysis, hate speech detection, and related socially grounded tasks, especially in settings where language-specific resources are limited.

Benchmarking Polarization. Shared evaluation campaigns have standardized socially relevant NLP tasks, including stance detection and misinformation analysis (Mohammad et al., 2018). Recent efforts extend this paradigm to multilingual polarization analysis by introducing large-scale benchmarks spanning diverse languages and sociopolitical contexts. These resources enable systematic study of multilingual robustness and cultural variation in polarized discourse.

Positioning of Our Work. Our work is positioned as a shared-task system paper centered on a simple and reproducible multilingual baseline. Rather than introducing a new architecture, we investigate the effectiveness of joint multilingual fine-tuning for binary polarization detection across a large set of languages, and we analyze where this setup succeeds and where it fails.

3 System Description

3.1 Dataset

We use the official multilingual corpus released for SemEval-2026 Task 9 (Naseem et al., 2026a), built upon the POLAR benchmark (Naseem et al., 2026b). The dataset contains user-generated social media posts annotated for binary polarization detection across 22 languages and diverse sociopolitical contexts.

Each instance includes an identifier, raw text, and a binary label. Training data are merged across languages for joint multilingual learning, while development and test sets are provided separately per language, supporting both multilingual and language-specific evaluation (Pires et al., 2019; Fortuna and Nunes, 2018).

3.2 Pre-processing

We adopt a minimal preprocessing strategy to preserve discourse-level cues relevant to polarized expression in social media text (Barbieri et al., 2020). Duplicate samples and instances with invalid labels are removed.

Text normalization is restricted to lowercasing when supported by the tokenizer. Punctuation, emojis, hashtags, and special characters are retained, as they often encode affective intensity and rhetorical framing associated with polarization.

Tokenization is performed using the SentencePiece tokenizer of XLM-RoBERTa (Conneau et al., 2020). Sequences are truncated or padded to a maximum length of 256 tokens, balancing contextual coverage and computational efficiency.

3.3 Model Architecture

Our system is built upon XLM-RoBERTa-large, a multilingual transformer pretrained with masked language modeling on large-scale cross-lingual corpora (Conneau et al., 2020). The model generates contextualized token representations for each input sequence.

Given an input sequence $x = (x_1, \ldots, x_n)$, the encoder produces hidden states $H = (h_1, \ldots, h_n)$. We use the representation of the special classification token, h_{cls}, as a sentence-level embedding. A linear classification layer is applied to predict the binary polarization label:

$$\hat{y} = \text{softmax}(W h_{\text{cls}} + b)$$

Here, W projects the sentence representation to a two-dimensional output space corresponding to the polarized and non-polarized classes, and the resulting logits are normalized with a softmax function. All model parameters are fine-tuned end-to-end.

3.4 Training Strategy

We adopt joint multilingual fine-tuning by merging all training data across languages into a single unified corpus. This setting allows the model to learn from a large and diverse multilingual training signal while maintaining a single shared classification pipeline across languages.

Optimization is performed using AdamW with a learning rate of 2×10^{-5}. Mixed-precision (FP16) training is employed to improve computational efficiency. We use a batch size of 16 with gradient accumulation to accommodate GPU memory constraints.

Model selection is based on Macro-F1 measured on the development set, and the checkpoint with the best development performance is retained for final evaluation.

4 Experimental Setup

4.1 Evaluation Measures

We evaluate system performance using the official metrics defined for the polarization detection setting. The primary metric is Macro-averaged F1 (Macro-F1), which assigns equal weight to each class and is appropriate for imbalanced and multilingual classification scenarios (Baccianella et al., 2009; Opitz and Burst, 2019).

In addition, we report Accuracy, Precision, Recall, and Micro-averaged F1 (Micro-F1) to provide complementary perspectives on model behavior, particularly for analyzing false positive and false negative tendencies in socially grounded tasks (Fortuna and Nunes, 2018).

For this binary setup, Macro-F1 is computed as:

$$F1_{\text{macro}} = \frac{1}{2}\left(F1_{\text{polarized}} + F1_{\text{non-polarized}}\right)$$

Micro-F1 aggregates prediction statistics across classes prior to F1 computation, emphasizing majority-class performance (Sokolova and Lapalme, 2009).

4.2 Implementation Details

Experiments were implemented in PyTorch using the Hugging Face Transformers library. Training was conducted on a single NVIDIA GPU with 48GB memory using mixed-precision (FP16) arithmetic. We used a learning rate of 2×10^{-5}, batch size 16, and gradient accumulation to increase the effective batch size under memory constraints. Model selection was based on development Macro-F1, and the best checkpoint was retained for final evaluation. The random seed was fixed for reproducibility. No language-specific preprocessing or language-dependent components were used.

Metric	Score
Average Accuracy	0.822
Average Macro-F1	0.780
Languages Evaluated	22

Table 2: Overall multilingual performance on the polarization detection task.

5 Results

5.1 Cross-Lingual Trends and Analysis

Table 2 presents the aggregated multilingual performance across all 22 languages. The overall Accuracy and Macro-F1 indicate that the jointly fine-tuned multilingual model provides competitive performance across typologically diverse languages and sociopolitical contexts.

Table 3 provides a detailed breakdown of performance per language. Macro-F1 scores exceed 0.75 for the majority of languages, demonstrating stable detection capability across both high-resource and moderate-resource settings. Languages such as Chinese, Hindi, Persian, and Bengali achieve particularly strong results, which is consistent with the overall robustness of the multilingual setup on several well-supported languages.

Notably, several lower-resource languages, including Amharic and Odia, also achieve competitive Macro-F1 scores under the joint multilingual setting. While this pattern is consistent with the hypothesis that multilingual training can support cross-lingual generalization, our experiments do not include a controlled comparison against language-specific fine-tuning, so this interpretation should be treated as suggestive rather than conclusive.

In contrast, languages such as Italian and Hausa exhibit lower Macro-F1 values. These degradations are primarily driven by Recall drops, indicating conservative prediction behavior and difficulties in identifying polarized content in linguistically sparse or culturally implicit contexts. Such trends highlight the ongoing challenges of modeling subtle ideological expressions in lower-resource environments.

Some language-specific metric combinations also indicate asymmetric class behavior. For example, Khmer shows very high Accuracy (0.920), Precision (0.926), and Recall (0.991), but a substantially lower Macro-F1 (0.650), suggesting that performance may be dominated by one class and that

Lang	Acc	Prec	Rec	M-F1
Amharic	0.835	0.847	0.949	0.759
Arabic	0.805	0.792	0.765	0.802
Bengali	0.839	0.809	0.810	0.835
German	0.718	0.716	0.681	0.717
English	0.794	0.757	0.647	0.771
Persian	0.848	0.873	0.930	0.788
Hausa	0.911	0.653	0.354	0.705
Hindi	0.905	0.929	0.962	0.796
Italian	0.665	0.768	0.419	0.639
Khmer	0.920	0.926	0.991	0.650
Burmese	0.857	0.868	0.885	0.854
Nepali	0.897	0.881	0.918	0.897
Odia	0.839	0.767	0.621	0.789
Punjabi	0.742	0.735	0.733	0.741
Polish	0.809	0.779	0.758	0.803
Russian	0.808	0.676	0.687	0.772
Spanish	0.768	0.772	0.752	0.768
Swahili	0.760	0.828	0.657	0.757
Telugu	0.866	0.948	0.784	0.866
Turkish	0.779	0.779	0.803	0.778
Urdu	0.816	0.839	0.910	0.771
Chinese	0.894	0.919	0.868	0.894

Table 3: Per-language performance across 22 languages (Acc, Prec, Rec, M-F1).

class-wise balance remains limited despite strong aggregate scores. A similar precision–recall imbalance is visible for Hausa, where high Accuracy (0.911) coexists with much lower Recall (0.354), indicating conservative prediction behavior. These patterns show why Macro-F1 is more informative than Accuracy alone for this task.

Overall, the per-language results show that joint multilingual fine-tuning provides a competitive and scalable baseline for the task, while also revealing instability in class balance for some languages. These findings motivate more systematic comparisons with language-specific adaptation and more detailed class-wise error analysis in future work.

5.2 Quantitative Error Analysis

A quantitative inspection of the per-language results reveals that the main source of degradation is not uniformly low performance, but unstable class balance across languages. In particular, some languages combine relatively high Accuracy with noticeably lower Macro-F1, indicating that one class is handled much better than the other. This behavior is especially visible in Khmer and Hausa, and to a lesser extent in Italian, where the precision–recall profile suggests conservative or skewed decision boundaries.

More specifically, Hausa obtains Accuracy 0.911 but Recall 0.354 and Macro-F1 0.705, which indicates that the system often misses polarized in-

stances despite strong aggregate correctness. Italian shows a similar tendency, with Precision 0.768 but Recall 0.419 and Macro-F1 0.639. In contrast, languages such as Nepali and Chinese exhibit more balanced performance, suggesting that the multilingual setup is more stable when the signal is either clearer or better aligned with the pretrained encoder.

5.3 Qualitative Error Analysis

A qualitative review of development-set errors suggests three recurrent failure modes. First, the model struggles with implicit polarization expressed through sarcasm or rhetorical inversion, where the literal wording is not overtly hostile but the intended stance is strongly divisive. Second, some false positives arise in posts that describe controversial political or social events in neutral terms; in these cases, the model appears to associate topic sensitivity with polarization. Third, culturally grounded references and locally meaningful expressions are sometimes misclassified because their polarizing function depends on shared background knowledge that is not explicit in the text itself.

These observations should be interpreted as qualitative tendencies rather than exhaustive error categories, since we did not annotate a large manual sample of false positives and false negatives. Nevertheless, they help explain why performance drops are concentrated in languages and contexts where pragmatic cues, implicit framing, and sociocultural knowledge play a central role.

From a practical perspective, the results indicate that a unified multilingual model can support large-scale polarization monitoring across diverse linguistic environments without requiring language-specific pipelines. At the same time, careful calibration is necessary to avoid over-detection in culturally sensitive contexts, where excessive false positives could distort downstream assessments of societal conflict.

Overall, the findings underscore both the scalability of multilingual transformer-based approaches and the continued need for culturally informed and discourse-level modeling to enhance robustness in complex real-world scenarios. A further implication of these error patterns is that multilingual performance should not be interpreted only through aggregate averages. In this task, languages with strong overall Accuracy may still exhibit unstable class-wise behavior, especially when polarization is expressed indirectly or through culturally specific framing. This suggests that future evaluations should place greater emphasis on per-language and class-sensitive analysis in order to better capture the real strengths and limitations of multilingual systems.

6 Conclusion

We presented DUTH, a joint multilingual fine-tuning approach for SemEval-2026 Task 9 Subtask 1 on binary polarization detection. Using a single XLM-RoBERTa encoder trained across 22 languages, the system achieved an average Accuracy of 0.822 and an average Macro-F1 of 0.780. These results show that a simple multilingual transformer provides a competitive and reproducible baseline for the task, while also revealing limitations in handling class imbalance, implicit polarization, sarcasm, and culturally specific references.

More broadly, these findings highlight the value of strong multilingual baselines in shared-task settings, where consistency, scalability, and reproducibility remain important alongside raw performance. Although the proposed system is intentionally simple, the analysis shows that it can still reveal meaningful cross-language trends and serve as a useful reference point for future language-adaptive approaches.

The per-language analysis further shows that strong aggregate performance does not always imply balanced class-wise behavior across languages. Future work will focus on controlled comparisons with language-specific adaptation, improved handling of class imbalance, and more systematic error analysis for implicit and culturally grounded cases.

References

Georgios Arampatzis and Avi Arampatzis. 2025. Hybrid sparse-neural fusion for passage retrieval. In *Proceedings of the Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST). RAGTIME Track.

Georgios Arampatzis, Ioannis Maslaris, and Avi Arampatzis. 2025a. LLM-based question generation and retrieval-augmented reporting for news credibility. In *Proceedings of the Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST). DRAGUN Track.

Georgios Arampatzis, Vasileios Perifanis, and Avi Arampatzis. 2025b. Justification retrieval with LLMs, retrieval-augmented generation, and hybrid labels. In *Proceedings of the Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST). RAG Track.

Georgios Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. 2025c. DUTH at EXIST 2025: Multilingual sexism detection with soft labels and transformers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9–12 September 2025*, CEUR Workshop Proceedings, pages 1793–1800. CEUR-WS.org.

Georgios Arampatzis, Konstantina Safouri, and Avi Arampatzis. 2025d. Bridging lexical and neural ranking for topic-oriented retrieval. In *Proceedings of the Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST). Tip-of-the-Tongue Track.

Giorgos Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. 2023. DUTH at SemEval-2023 task 9: An ensemble approach for Twitter intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1225–1230, Toronto, Canada. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 283–287.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

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.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

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.

Juri Opitz and Sebastian Burst. 2019. Macro F1 and macro F1. *arXiv preprint arXiv:1911.03347*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437.

Cass R. Sunstein. 2009. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.