POLITECNICO DI TORINO
Repository ISTITUZIONALE

MINDS at SemEval-2026 Task 9: A Multi-Paradigm Approach to Cross-Lingual Polarization Detection

Original
MINDS at SemEval-2026 Task 9: A Multi-Paradigm Approach to Cross-Lingual Polarization Detection / Iannielli, A., Maroli, S., Roberto, M., Sammartino, S., Vacirca, V., Savelli, C., Coppola, R., Giobergia, F.. - (2026), pp. 2475-2486. (20th International Workshop on Semantic Evaluation (SemEval-2026) San Diego, CA (USA) July 2–7, 2026) [10.18653/v1/2026.semeval-1.313].

Availability:
This version is available at: 11583/3015347 since: 2026-09-10T16:08:39Z

Publisher:
Association for Computational Linguistics

Published
DOI:10.18653/v1/2026.semeval-1.313

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

25 November 2026

MINDS at SemEval-2026 Task 9: A Multi-Paradigm Approach to Cross-Lingual Polarization Detection

Angelo Iannielli* **Samuele Maroli*** **Marco Roberto*** **Stefano Sammartino***
Valentino Vacirca* **Claudio Savelli†** **Riccardo Coppola†** **Flavio Giobergia†**
Politecnico di Torino
* {firstname.lastname}@studenti.polito.it
† {firstname.lastname}@polito.it

Abstract

Online polarization has become a central challenge in digital discourse, characterized by hostility, identity-based division, and culturally dependent expressions that vary across languages. Automatically detecting such phenomena is particularly difficult in multilingual settings, where semantic nuance and implicit rhetoric complicate cross-lingual generalization. In this context, we participate in POLAR, a shared task at SemEval 2026 on multilingual polarization detection and categorization across 22 languages. We compare three modeling paradigms: multilingual encoder fine-tuning, translation-based transfer learning, and prompting-based generative reasoning. For the multi-label categorization task, we introduce a two-stage cascaded architecture to mitigate false positives under severe class imbalance. Our results show that multilingual encoders achieve the most robust performance for binary detection, whereas reasoning-based prompting is competitive for fine-grained category classification. This comparative study highlights the strengths and limitations of each paradigm for cross-lingual polarization analysis.

1 Introduction

Online polarization has emerged as a critical challenge in modern digital discourse, characterized not merely by disagreement but by sharp division, hostility, and "us-vs-them" dynamics between identity groups. While this phenomenon is a precursor to social fragmentation and radicalization, automated detection systems have primarily focused on high-resource languages such as English. Extending these capabilities to a global scale remains an open challenge: implicit cultural nuances, varying linguistic typologies, and domain-specific contexts make simple transfer learning insufficient. As noted in the cross-lingual literature (Conneau et al., 2018), models trained on a single language often fail to capture the semantic shifts that occur when discourse crosses cultural boundaries.

In this paper, we address Task 9 (Naseem et al., 2026a) at SemEval 2026. The shared task is based on the POLAR benchmark (Naseem et al., 2026b) and addresses the complexity of polarization across 22 languages, ranging from high-resource languages such as Spanish to low-resource languages such as Amharic and Burmese. The experimental setting poses a dual challenge: massive linguistic diversity and severe class imbalance. It is unclear whether the "strong baseline" of translation-based approaches can preserve the subtle semantics of polarization, or if native multilingual encoders offer superior representations.

To address this uncertainty, we conducted a Comparative Multi-Paradigm Analysis[1]. Rather than proposing a monolithic system, our goal was to determine the most effective strategy for cross-lingual polarization detection by comparing three fundamentally different architectural philosophies:

1. **Encoder-only Approach (EA):** We employed a parameter-efficient multilingual encoder (XLM-RoBERTa (Conneau et al., 2020) with LoRA (Hu et al., 2022)) to directly model cross-lingual representations. For Subtask 2, we further introduced a Two-Stage Cascaded Pipeline to mitigate class imbalance and reduce false positives in multi-label classification.

2. **Translation-based Approach (TA):** Following a *Translate-Test* paradigm inspired by Conneau et al. (2018), we mapped all texts into English using MarianMT (Junczys-Dowmunt et al., 2018) and trained a strong monolingual classifier (DeBERTa v3 (He et al., 2021)) on the translated data.

[1]The code to replicate the experiments can be found at https://github.com/MarcoRob919/MINDS-SemEval26-Polar

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2475–2486
July 3-4, 2026 ©2026 Association for Computational Linguistics

3. **Prompting-based Approach (PA):** We evaluated a large language model (Llama 3.1 8B q4bit (Meta AI, 2024)) without task-specific fine-tuning, leveraging structured prompting and explicit rationale generation, a strategy we refer to as *generative knowledge augmentation*.

2 POLAR task

The POLAR shared task (Naseem et al., 2026a) focuses on detecting and characterizing online polarization across 22 languages, addressing both high- and low-resource settings. The primary objective of this competition is to introduce a framework for the automatic identification of multilingual, multicultural, and multievent polarization. The competition comprises three distinct subtasks designed to model these complementary aspects. In this work, we focus exclusively on Subtasks 1 and 2.

1. **Subtask 1: Binary Polarization Detection** It requires a binary classification to determine whether a text contains polarized opinion. A text is labeled *Polarized* only if it clearly reflects attitude polarization (e.g., hostility, division) rather than just negative sentiment.

2. **Subtask 2: Polarization Type** This is a multi-label classification task to identify the target or domain of the polarization. The categories are *Political/Ideological*, *Racial/Ethnic*, *Religious*, *Gender/Sexual Orientation*, and *Other* (e.g., economic or media-based targets).

3. **Subtask 3: Manifestation Identification** This fine-grained multi-label task focuses on identifying specific rhetorical strategies that express polarization, such as *Vilification*, *Stereotyping*, and *Dehumanization*.

The experimental dataset comprises $73,681$ labeled instances across 22 languages. Globally, the first classification task exhibits a fairly balanced distribution, with approximately 53% of samples classified as polarized and 47% as non-polarized. A granular analysis has revealed significant variance in polarization rates across individual languages. For instance, *Hausa* exhibits a polarization rate of only $\sim 11\%$, whereas *Khmer* exceeds 90%. Regarding Subtask 2, the dataset consists of a subset of $66,312$ instances. Unlike the binary classification task, this multi-label classification challenge exhibits notable class imbalance. Political/Ideological polarization is the predominant category, vastly outnumbering others, whereas Gender/Sexual polarization accounts for less than 9% of the labeled data. This skew posed significant challenges during the training phase, necessitating specific mitigation strategies that we discuss in Section 3.1. Detailed statistics are provided in Table 1.

Subtask 1	**Count**
Total samples	$73,681$
Languages covered	22
Polarized	$39,145$ (53%)
Non-polarized	$34,536$ (47%)
Subtask 2	
Total samples	$66,312$
Political	$20,184$ (27%)
Other	$13,702$ (19%)
Racial/Ethnic	$11,724$ (16%)
Religious	$7,564$ (10%)
Gender/Sexual	$6,252$ (8.5%)

Table 1: Dataset statistics

3 Proposed Methodologies

In this section, we describe the three strategies evaluated for multilingual polarization detection.

3.1 Encoder-only Approach (EA)

Our first architectural approach leverages an encoder-based masked language model pre-trained on data in different languages, making it highly effective for cross-lingual transfer learning.

For Subtask 1 (binary detection), we attached a standard classification head to the encoder and fine-tuned the model using a cross-entropy loss. Although the problem is fairly well-balanced, we apply class weights inversely proportional to class frequencies to slightly improve the representation of the minority class. For Subtask 2 (multi-label classification), initial experiments with a single multi-label model revealed a critical flaw: the model frequently predicted polarization types for non-polarized texts. To address this problem, we implemented a Two-Stage Cascaded Pipeline, as shown in Figure 1, where Stage 1 acts as a filter, where the same binary classifier used in Subtask 1 predicts whether the text contains polarized content; and Stage 2 performs the classification task only if Stage 1 predicts a positive outcome. Otherwise, the system predicts no class.

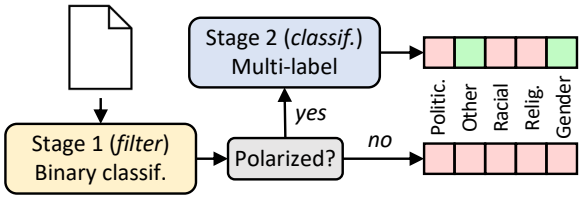

Figure 1: The Two-Stage Cascade Pipeline architecture. Non-polarized texts are filtered out at Stage 1 to prevent false label attribution in Subtask 2.

This architecture ensures that the Stage 2 model addresses only cases where polarization occurs, avoiding neutral content. The Stage 2 model was trained using Binary Cross-Entropy (BCE) loss.

3.2 Translation-based Approach (TA)

To address the dataset's multilingual nature, we also developed a translation-based strategy. This setting corresponds to a translate-train pipeline, where all training and evaluation texts are translated into a single target language (English). Previous work has shown that, when machine translation quality is sufficiently high, translate-train approaches can achieve performance comparable to multilingual models while using a strong monolingual encoder (Conneau et al., 2018). This design choice allowed us to unify the data into a single dataset, enabling us to use a strong English pretrained language model without relying on multilingual representations.

3.3 Prompting-based Approach (PA)

As an alternative to the approaches presented so far, we also explored a strategy using a pre-trained large language model. To effectively leverage the model's generative capabilities, we introduced an explicit reasoning step: the model must provide a brief explanation of why a given comment is polarized or non-polarized, and subsequently output a set of boolean flags representing the final classification. The inclusion of reasoning examples (rationales) helped generate explanations that were less trivial and more explicitly grounded in the lexical content of the input comment. To obtain easily processable outputs, we defined a structured JSON-style output format. The interaction with the model follows a few-shot approach, providing a set of illustrative examples that guide the model in producing well-structured outputs. We provide an example of an input/structured output pair in Figure 2.

The final output consists of a textual rationale, a binary flag indicating whether the comment is polarized, and a set of flags corresponding to each polarization category. The richness of this output allows the same architecture to be effectively employed for both subtasks. We translate all texts into English also for this solution. In addition, we identified edge cases in which the model predicts the absence of polarization content while still classifying some polarization classes as present. In these cases, we apply a post-processing step that ensures no polarization class is predicted when the "presence" outcome is negative.

Example of Rationale + Classification

Input Comment:
"Shut up little girl you didn't even know where she was."

LLaMA 3.1 8B q4bit Output:

```
{
"rationale": "The comment uses
derogatory language ('little girl') to
address someone, expressing hostility
and disrespect.",
"isPolarized": true,
"political": false,
"racial_ethnic": false,
"religious": false,
"gender_sexual": true,
"other": false
}
```

Model Prediction: Polarized (Category: gender/sexual)

Figure 2: Qualitative example of Knowledge Augmentation. The LLM explicits the latent hostility in a grammatically neutral sentence, bridging the cultural gap often missed by translation-based baselines.

4 Experimental Setup

For the encoder-only approach, we used XLM-RoBERTa (XLM-R) Large with LoRA for parameter-efficient fine-tuning. In preliminary experiments, XLM-R Large provided only marginal improvements over the Base variant. Nevertheless, we retained the Large model for the final submission in order to be more competitive in the challenge. To validate our PEFT strategy, we also compared LoRA with full fine-tuning to verify whether updating a limited subset of parameters would reduce the model's ability to capture complex polarization patterns.

For the translation-based approach, we adopted

MarianMT, an efficient neural machine translation model, to translate the non-English training data into English. These translated sentences were then merged with the original English data to construct a monolingual training set. For downstream classification, we used DeBERTa v3 Base fine-tuned on the aggregated dataset. To evaluate cross-lingual generalization, we reported, for computational reasons, the performance on a subset of 12 languages: English plus 11 languages translated into English. While automatic machine translation inherently introduces potential noise and may distort culturally specific nuances, it largely preserves the core semantic structures necessary for classification.

For the prompting-based approach, we used LLaMA 3.1 8B Instruct. This instruction-tuned model was applied without task-specific fine-tuning and was prompted to produce structured predictions for the target labels accompanied by explicit reasoning. This configuration enabled us to assess the baseline efficacy of modern instruction-following language models in resolving the task strictly through prompt engineering and in-context learning.

All experiments were conducted on Google Colab Pro using an NVIDIA T4 GPU with 15 GB of VRAM.

5 Results

All tables in this section report the F1-score metric, computed as the arithmetic mean across the 12 languages used for TA and PA for comparison. The text also reports the overall results for all languages for EA.

5.1 Subtask 1 results

Table 3 shows the results of all three architectures proposed for Subtask 1. The results clearly show better performance for EA when compared against the other pipelines (TA, PA).

The comparative experiment using full fine-tuning and LoRA on XLM-R Base showed similar outcomes ($F1_{macro} = 0.7666$ with full fine-tuning and $F1_{macro} = 0.7698$ with LoRA). Given the marginal overall difference and LoRA's slightly superior performance, we opted to implement the latter. Afterward, our experiments with XLM-R Base showed diminishing returns when increasing model capacity through LoRA adaptation. Specifically, raising the LoRA rank from 16 to 32 did not improve performance, with $F1_{macro}$ varying only marginally (from 0.7805 to 0.7702). This indicates that a lower rank is sufficient for the binary detection task. Although the XLM-R Large model achieved the highest absolute scores ($F1_{macro}$ of 0.7934 across all 22 languages), the improvement over the Base variant was limited. For completeness and competitiveness in the shared-task setting, Table 3 reports results obtained with XLM-R Large (EA) and LoRA with rank 32. Nevertheless, the Base model with rank 16 remains a more computationally efficient solution, delivering competitive performance while requiring fewer resources.

For the translation-based method (TA), we conducted a limited hyperparameters tuning study by varying the learning rate and max length, but

	Subtask 1		Subtask 2	
Lang.	**MINDS**	**baseline**	**MINDS**	**baseline**
amh	**0.7730**	0.7151	-	**0.3716**
arb	**0.8181**	0.7957	0.4649	**0.4855**
ben	0.8219	**0.8528**	0.2059	**0.2887**
deu	**0.7067**	0.6714	**0.4942**	0.4078
eng	**0.7943**	0.7802	-	**0.3333**
fas	0.8105	**0.8424**	-	**0.4626**
hau	0.7641	**0.7753**	-	**0.2038**
hin	**0.7959**	0.7379	0.2987	**0.7911**
ita	0.5992	**0.6773**	**0.4692**	0.3759
khm	**0.7139**	0.6592	-	**0.6268**
mya	**0.8558**	0.8210	-	**0.4772**
nep	**0.9002**	0.8798	-	**0.7219**
ori	**0.7778**	0.7765	-	**0.5600**
pan	0.7586	**0.7898**	-	**0.3650**
pol	**0.7953**	0.7241	0.4207	**0.4491**
rus	**0.7997**	0.7457	0.3911	**0.5904**
spa	**0.7827**	0.7266	0.4723	**0.5935**
swa	**0.7718**	0.7571	-	**0.4417**
tel	**0.8677**	0.6440	-	**0.3145**
tur	**0.7792**	0.6957	**0.4708**	**0.4708**
urd	**0.7918**	0.7890	0.2515	**0.7127**
zho	**0.8836**	0.8691	0.5086	**0.6697**

Table 2: Comparison between MINDS and baseline across Subtasks 1 and 2. Best score per language and subtask in bold.

Method	**F1$_{\text{macro}}$**
EA (XLM-RoBERTa Large)	0.7930
TA (MarianMT + DeBERTa)	0.7060
PA (LLaMA Instruct + Reasoning)	0.6555

Table 3: Results of Subtask 1 over the 12 languages used for TA and PA for comparison across approaches.

these experiments did not yield substantial improvements; therefore, we retained the configuration with learning rate $2 \cdot 10^{-5}$ and max length $= 384$, which achieved the best score ($F1_{macro} = 0.7060$). However, the language-wise evaluation shows noticeable variability: English achieved the highest score ($F1_{macro} = 0.7920$), while Italian ($F1_{macro} = 0.5983$) and Urdu ($F1_{macro} = 0.6011$) exhibited the largest drops. This dispersion is plausibly explained by translation-induced variability, where differences in translation quality, ambiguity resolution, and the handling of idiomatic or culturally specific expressions may alter task-relevant nuances and reduce classification consistency across languages.

5.2 Subtask 2 results

Table 4 shows the results of all three architectures proposed for Subtask 2.

Method	$\mathbf{F1_{macro}}$
EA (Two-Stage Cascade Pipeline)	0.5630
TA (MarianMT + DeBERTa)	0.4842
PA (LLaMA Instruct + Reasoning)	0.6335

Table 4: Results of Subtask 2 over the 12 languages used for TA and PA for comparison across approaches.

For the encoder-only approach (EA), we conducted a series of hyperparameter-tuning experiments to improve performance. We tested configurations that increased dropout, optimized class thresholds, and targeted both the Attention and FFN layers for LoRA adaptation. The highest result across all languages was $F1_{macro}$, with a value of 0.6524. The chosen hyperparameters are listed in Table 7 in the Appendix. In contrast, the Cascade strategy exhibited distinct behavior: although accuracy was lower than that of the optimized single model, the pipeline achieved a significantly higher $F1_{macro}$ equal to 0.6946 over all 22 languages, suggesting that a serialized architecture is more effective at capturing minority-label instances than a heavily tuned single large model.

For the translation-based approach (TA), we followed the same hyperparameter tuning strategy described for Subtask 1, varying the learning rate and maximum sequence length. As in the binary setting, these experiments did not lead to substantial performance gains. Consequently, we retained the configuration with learning rate $2 \cdot 10^{-5}$ and max length $= 384$, which achieved the best overall result ($F1_{macro} = 0.4842$). Language-wise performance in Subtask 2 differs from that observed in Subtask 1. Although the same hyperparameter configuration is retained, Urdu achieves the highest score, while Bengali shows the lowest performance, as visible in the results presented in Appendix B. This variability underscores the non-uniform effects of translation across languages, particularly in a multi-label setting, where subtle semantic shifts can affect category-specific predictions.

Overall, the prompting-based approach with explicit reasoning achieves the highest macro F1 score among all evaluated methods for Subtask 2. This suggests that fine-grained polarization type classification benefits from the richer semantic modeling enabled by generative reasoning. Unlike encoder-based architectures, which rely solely on latent representations, the reasoning-augmented LLM appears better suited to capture subtle target-specific cues across languages.

5.3 Inclusion of reasoning

To assess the usefulness of the reasoning process in Subtasks 1 and 2, we conducted two experiments with the PA approach. In the first setting, the model was required to generate a textual rationale alongside its predictions; in the second, it produced only the classification labels. The prompt structure and few-shot examples were kept identical across both settings to ensure a fair comparison. As reported in Table 5, requiring a reasoning step consistently improves performance.

PA Setting	ST 1 ($\mathbf{F1_{macro}}$)	ST 2 ($\mathbf{F1_{macro}}$)
No Reasoning	0.5999	0.5963
With Reasoning	**0.6555**	**0.6335**

Table 5: Macro F1 performance for PA on Subtasks 1 and 2 (ST1, ST2) with/without reasoning.

It is interesting noting that, although rarely, the model may produce structurally invalid outputs. For instance, it may refuse to respond to particularly strong comments or fail to generate a complete JSON object. This issue was significantly mitigated by enriching the prompt context, adopting the official chat template, and providing more detailed, restrictive instructions in the system section. Across the entire final validation process, in which both approaches (with and without reasoning) were executed, only a single invalid output was observed.

6 Analysis and Discussion

In this section, we analyze model behavior and limitations. We first examine the impact of the cascaded architecture, then discuss common error sources, and finally interpret the results through the lens of implicit and explicit polarization.

6.1 Impact of the Cascaded Architecture

The introduction of a cascade system significantly reduced false positives in the polarization classifier. The intermediate reasoning step helped prevent topic-triggered activations of sensitive labels when the comment itself was not polarized. Three representative German examples from the final test set are shown in Appendix C: in each case, the initial model incorrectly assigned polarization-related labels due to the presence of politically salient topics, while the cascade system correctly identified the comments as non-polarized. These cases illustrate how explicit reasoning mitigates over-activation of sensitive categories and improves robustness in scenarios where topic intensity does not correspond to actual polarized intent.

6.2 Translation-Induced Errors

Using a lightweight model such as MarianMT reduced preprocessing costs but introduced substantial translation noise. Manual inspection revealed frequent hallucinations (e.g., repeated tokens, untranslated colloquialisms or fully fabricated sequences), a phenomenon also observed in generative language models (Borra et al., 2024), as well as occasional meaning inversion and failures triggered by noisy input. Some errors preserved classification-relevant cues, while others replaced the original semantics entirely. These issues were more pronounced in low-resource languages such as Urdu, where translations were often incorrect or semantically unrelated to the source text, likely due to limited representation in the model's training data (see Appendix D for examples).

6.3 Implicit vs. Explicit Polarization: A Qualitative Analysis

A useful perspective for interpreting our results is the distinction between *explicit* and *implicit* polarization. Explicit cases convey hostility through surface lexical cues (slurs, overt insults) and are largely recoverable from word-level signals. Implicit cases instead rely on sarcasm, stereotypes, assumptions, or culturally loaded framings, and can appear lexically neutral while still carrying strong identity-based hostility. The example in Figure 2 illustrates this: the patronizing use of "little girl" encodes gendered hostility without any explicit marker.

This distinction helps explain the relative behavior of our three paradigms. EA excels on Subtask 1, where binary detection is often resolvable from distributional lexical patterns typical of explicit polarization. PA, in contrast, overtakes EA on Subtask 2, where identifying the targeted group frequently requires pragmatic and world-knowledge reasoning over implicit cues. The weakness of TA can be read along the same axis: as our error analysis shows (Appendix D), translation tends to preserve explicit content while degrading idioms, sarcasm, and culture-bound references. This loss is amplified in low-resource languages such as Urdu, where translation noise disproportionately removes the pragmatic layer needed to detect slight polarization.

7 Conclusion

In this paper, we presented a comparative study of three complementary strategies for multilingual polarization analysis in the context of the POLAR shared task. We evaluated native multilingual encoding with parameter-efficient fine-tuning, a translation-based transfer approach, and a prompting-based generative strategy with explicit reasoning, focusing on Subtasks 1 and 2.

Our results show that a fine-tuned multilingual encoder provides the most robust solution for binary polarization detection, achieving strong and stable performance across languages. For the multi-label categorization task, the prompting-based approach with explicit reasoning demonstrated competitive performance, suggesting that fine-grained target identification benefits from richer semantic modeling. In contrast, the translation-based pipeline consistently underperformed the other approaches, suggesting potential limitations when subtle cultural and stylistic cues are involved.

Overall, our findings indicate that distinct subtasks in multilingual polarization analysis may benefit from different modeling paradigms. Future work will explore hybrid architectures that combine efficient multilingual encoders with reasoning-aware training strategies, aiming to integrate classification accuracy and semantic interpretability within a unified framework.

References

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684.

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.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R. Bowman. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, and 1 others. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Meta AI. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

A Hyperparameter Configurations

Hyperparameter	Value
Fine-tuning method	LoRA
LoRA rank (r)	32
LoRA alpha (α)	64
LoRA dropout	0.05
Max sequence length	384
Training batch size	16
Evaluation batch size	32
Gradient accumulation steps	2
Learning rate	1.5×10^{-4}
Number of epochs	5
Warmup ratio	0.1
Weight decay	0.01
Optimizer	AdamW
Learning-rate scheduler	Cosine
Loss function	Weighted cross-entropy
Random seed	42

Table 6: Hyperparameters used for the XLM-RoBERTa Large model with LoRA in Subtask 1.

Hyperparameter	Value
Shared Configuration	
Fine-tuning method	LoRA
Max sequence length	256
Evaluation batch size	32
Warmup ratio	0.10
Weight decay	0.01
Gradient accumulation steps	2
Random seed	42
Stage 1: Binary Polarization Filter	
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.10
Learning rate	2×10^{-4}
Number of epochs	5
Training batch size	32
Stage 2: Multi-label Type Classifier	
LoRA rank (r)	32
LoRA alpha (α)	64
LoRA dropout	0.05
Learning rate	1.5×10^{-4}
Number of epochs	5
Training batch size	32

Table 7: Hyperparameter configuration of the two-stage cascaded pipeline used for Subtask 2. Both stages share the same XLM-RoBERTa Base backbone, with task-specific LoRA configurations.

Hyperparameter	Value
Fine-tuning method	Full fine-tuning (no LoRA)
Max sequence length	384
Training batch size	16
Evaluation batch size	32
Learning rate	2×10^{-5}
Number of epochs	5
Warmup ratio	0.06
Weight decay	0.01
Gradient accumulation steps	1
Optimizer	AdamW
Learning-rate scheduler	Cosine
Loss function	Weighted cross-entropy
Early stopping patience	2
Random seed	42

Table 8: Hyperparameters used for the DeBERTa v3 Base model in Subtask 1.

Hyperparameter	Value
Fine-tuning method	Full fine-tuning (no LoRA)
Max sequence length	384
Training batch size	16
Evaluation batch size	32
Learning rate	2×10^{-5}
Number of epochs	3
Warmup ratio	0.06
Weight decay	0.01
Gradient accumulation steps	1
Optimizer	AdamW
Learning-rate scheduler	Cosine
Loss function	BCEWithLogitsLoss
Random seed	42

Table 9: Hyperparameters used for the DeBERTa v3 Base model in Subtask 2.

B Per language results

Language	EA (XLM-RoBERTa Large)	TA (MarianMT + DeBERTa Base)	PA (LLaMA Instruct + Reasoning)
arb	**0.8178**	0.7613	0.7479
ben	**0.8439**	0.7436	0.7173
deu	**0.7339**	0.6632	0.7128
en	0.7788	**0.7920**	0.7080
hin	**0.7802**	0.6504	0.5938
ita	**0.6317**	0.5983	0.3877
pol	**0.7882**	0.7513	0.6653
rus	**0.7872**	0.7489	0.6154
spa	**0.7295**	0.7158	0.6749
tur	**0.8369**	0.6736	0.7043
urd	**0.7515**	0.6111	0.5421
zho	**0.8887**	0.7626	0.7966

Table 10: $F1_{macro}$ per language for Subtask 1.

Language	EA (XLM-RoBERTa Large)	TA (MarianMT + DeBERTa Base)	PA (LLaMA Instruct + Reasoning)
arb	0.5750	0.5245	**0.6904**
ben	0.4223	0.2946	**0.5690**
deu	0.4892	0.5251	**0.6964**
en	0.4554	0.4140	**0.6255**
hin	**0.7223**	0.5145	0.5594
ita	0.2719	0.3211	**0.5482**
pol	0.4920	0.4765	**0.6805**
rus	0.5693	0.4575	**0.6625**
spa	0.6391	0.5825	**0.7183**
tur	0.6606	0.4505	**0.6831**
urd	**0.7552**	0.6989	0.4296
zho	**0.7688**	0.5502	0.7393

Table 11: $F1_{macro}$ per language for Subtask 2.

Language	Subtask 1 (EA)	Subtask 2 (PA)
arb	0.8181	0.4649
ben	0.8219	0.2059
deu	0.7067	0.4942
en	0.7943	0.6255
hin	0.7959	0.2987
ita	0.5992	**0.4692**
pol	0.7953	0.4207
rus	**0.7997**	0.3911
spa	**0.7827**	0.4723
tur	0.7792	0.4708
urd	**0.7918**	0.2515
zho	0.8836	0.5086

Table 12: $F1_{macro}$ results of the challenge. The scores in the top 10 of the rankings are highlighted in bold.

C Examples of cascade contribution (PA)

ID	Translation (EN)	Classification flags (before applying the cascade)
deu_ e90a8acf 58d89d44 2212a208 341c8775	*Let's see what Ukrainian law says.*	**isPolarized: false** **political: true** racial_ethnic: false religious: false gender_sexual: false other: false
deu_ f7d1a003 c9251d7c 4e97bd52 dc1b293c	*There were also reports of freed Hamas hostages that they were taken hostage by civilians and then sold to Hamas...*	**isPolarized: false** **political: true** **racial_ethnic: true** **religious: true** gender_sexual: false other: false
deu_ 2cf78df1 d2eaaef0 f71c0eca a75fabf8	*Has the hatred of FFF passed by you? ...*	**isPolarized: false** **political: true** racial_ethnic: false religious: false gender_sexual: false other: false

Table 13: Examples of false positives mitigated by the cascade reasoning system.

D Examples of translation errors

ID	Source text	Model translation	Error type / Notes
ita_f08419ab79f26aa91429125dc81ed343	*"VI STA' BENE! [...] DEVONO TOGLIERVI TUTTO."*	Long hallucinated repetition of "it's okay"	Hallucination, meaning lost.
ita_6dae411033309d164450a6d7f921959f	*"Insuperabile Gabriella [...] no comment"*	Correct first part + long sequence of "."	Token repetition hallucination.
ita_a01d8cc950d6d021a3a703e4a40f7bd0	*"[...] Un pirla!"*	*"[...] a pirla!"*	Untranslated colloquialism.
ita_584c2ccefc4d2fd73cf0ab8f91c2343b	*"È il colpevole."*	*"It's the killer."*	Semantic shift (a "colpevole" is not necessarily a "killer").
ita_cb4c33be2240ef71797ecc911b3a4009	*"DACCORDISSIMO!!!!!!"*	*"DACCORDISSIMO!!!!!!"*	Typo-induced failure.
spa_937f9818c972b14688234e53f7894a04	*"la zorra en el gallinero"*	*"The bitch in the henhouse"*	Wrong sense disambiguation ("zorra" = fox).
urd_12a59fe6550a4dbd84c07734244ce978	*Sentence about fatwa and legal action*	Nonsensical sequence ("right-to-do-do-to-do…")	Catastrophic hallucination; meaning lost.
urd_ec1c4f3e841ec54331a5f824122e4572	*Sentence about media and establishment*	Output of repeated "A_A_A..." tokens	Severe corruption; no semantic content.

Table 14: Examples of translation errors across Italian, Spanish, and Urdu comments.