A Multi-Paradigm Approach for Cross-Lingual Polarization Analysis

Iannielli Angelo ID 317887 | **Maroli Samuele** ID 334151 | **Roberto Marco** ID 339387 | **Sammartino Stefano** ID 333676 | **Vacirca Valentino** ID 346945

Abstract

Polarization analysis in social media comments is challenging due to the diversity of topics, languages, and forms of implicit hostility. We compare three complementary approaches (multilingual encoding, translation-based transfer, and reasoning prompting) across two sub-tasks: polarization detection and category assignment. Our study includes a two-stage cascaded pipeline and a translation-based fine-tuning setup evaluated both globally and per language. We also test a non-fine-tuned generative model using explicit reasoning to justify predictions. This multi-paradigm comparison highlights the strengths and limitations of each strategy for cross-lingual polarization analysis.

GitHub repository

The source code for the project is available on GitHub at the following address:
`https://sl1nk.com/github-ValTal5-LLM4SE`

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 serves as a precursor to social fragmentation and radicalization, automated detection systems have predominantly focused on high-resource languages like 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 cross-lingual literature (Conneau et al., 2018), models trained on single languages often fail to capture the semantic shifts that occur when discourse crosses cultural boundaries.

This report describes our team's participation in `Task 9 of POLAR @ SemEval 2026`. The shared task (Usman Naseem, 2026) addresses the complexity of polarization across **22 languages**, ranging from high-resource ones like Spanish to low-resource ones like Amharic and Burmese. The experimental setting presents a dual challenge: a massive linguistic diversity and a severe class imbalance. In this context, relying on a single architectural philosophy is risky. 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**. Rather than proposing a monolithic system, we evaluated the trade-offs between three distinct strategies: *native multilingual encoding* (to capture latent alignment), *translation-based transfer* (to leverage monolingual strength), and *generative knowledge augmentation* (to explicitly model reasoning).

1.1 Task Formulation and Scope

The competition is structured into three distinct subtasks designed to capture different granularities of polarized discourse.

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 used to express polarization, such as *Vilification*, *Stereotyping*, or *Dehumanization*.

1.2 Research Objectives and Comparative Approach

Our primary goal was to determine the most effective strategy for cross-lingual polarization detection by comparing three fundamentally different architectural philosophies:

1. **Native Encoding vs. Translation:** We investigated whether mapping all languages to English outperforms native multilingual representations. Following the methodology of (Conneau et al., 2018), we compared a *Translate-Test* baseline (using **MarianMT** (Junczys-Dowmunt et al., 2018) + **DeBERTa v3** (He et al., 2021)) against a parameter-efficient native encoder (**XLM-RoBERTa** (Conneau et al., 2020) with LoRA (Hu et al., 2022).

2. **Latent vs. Explicit Reasoning:** Standard encoders rely on latent vector representations which may miss cultural context. We analyzed if explicitly extracting "rationales" via a Large Language Model, **Llama 3.1 8B q4bit** (Meta AI, 2024), a strategy of Generative Knowledge Augmentation, improves classification performance compared to standard discriminative baselines.

3. **Handling Imbalance:** We evaluated the impact of a **Two-Stage Cascaded Pipeline** for Subtask 2, assessing its ability to reduce false positives in underrepresented categories compared to direct multi-label classification.

The paper is organized as follows. Section 2 provides the necessary background and discusses related work in cross-lingual polarization. Section 3 presents a quantitative analysis of the dataset, highlighting the linguistic disparities and class imbalance that motivated our choices. Section 4 details the architectural implementation of our three paradigms: native multilingual encoding, translation-based transfer, and generative augmentation. Finally, Section 5 discusses the comparative experimental results, followed by conclusions and future work in Section 6.

2 Background

2.1 Online Polarization and NLP

Online polarization refers to language that expresses division, hostility, or opposition toward specific social groups, ideologies, or identities. In contrast to general sentiment analysis, polarization is often group-oriented and may target political actors, ethnic or religious communities, or gender-related identities. For this reason, polarization detection is closely related to tasks such as hate speech and bias detection, but it also includes more subtle and implicit forms of antagonism.

From an NLP perspective, detecting polarization is challenging for several reasons. Polarized content is often context-dependent and may not rely on explicit offensive terms. Moreover, a single text may express polarization toward multiple targets at the same time, making the task inherently multi-label. These challenges are further amplified in multilingual settings, where the same polarized attitude can be expressed differently across languages and cultures.

2.2 Multilingual and Cross-lingual Approaches

Previous works has addressed multilingual text classification using two main strategies. The first relies on multilingual pretrained models, such as XLM-RoBERTa, which are trained on data from many languages and can be fine-tuned jointly on multilingual datasets. The second strategy adopts a translation-based approach, where all texts are translated into a single high-resource language, typically English, and a monolingual model is trained on the translated data. Both approaches have been shown to be effective in cross-lingual classification tasks.

2.3 Parameter-efficient Fine-Tuning

More recently, the growing size of pretrained language models has led to increased interest in parameter-efficient fine-tuning methods. Techniques such as **Low-Rank Adaptation (LoRA)** allow models to be adapted to new tasks by training only a small number of additional parameters, reducing computational cost while maintaining strong performance. These methods are particularly useful in multilingual scenarios, where training full models can be expensive.

2.4 Prompting-based Approaches with Large Language Models

Recent advances in large language models (LLMs) have enabled a different approach to text classification based on prompting rather than fine-tuning. Instruction-following models, such as LLaMA and similar decoder-only architectures, can perform complex reasoning tasks by conditioning on carefully designed prompts and a small number of in-context examples.

In this paradigm, the model is guided to reason explicitly about the input text before producing a final prediction, often following a chain-of-thought style. This approach has been shown to be effective in tasks requiring nuanced interpretation, such as stance detection, bias analysis, and content moderation. However, prompting-based methods may produce verbose or inconsistent outputs, making post-processing steps necessary to extract structured predictions and enforce logical constraints.

Despite these limitations, prompting-based approaches provide a flexible alternative to supervised fine-tuning, especially when labeled data is limited or when interpretability of the model's reasoning is desired.

3 Explorative Data Analysis

The experimental dataset comprises a total of $73,681$ labeled instances distributed 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 task, this multi-label classification challenge presents a 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 4.1.2.

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

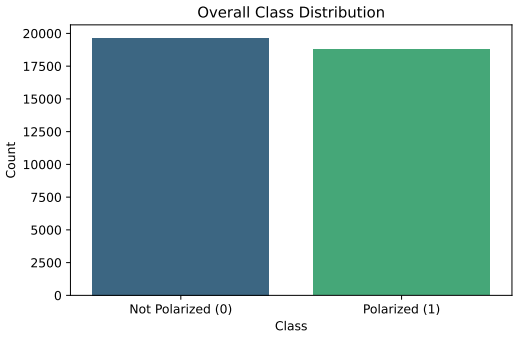

Figure 1: Distribution of polarization in Subtask 1.

4 System overview

4.1 Parameter-Efficient XLM-RoBERTa

Our primary architectural approach leverages the **XLM-RoBERTa** (XLM-R), a transformer-based masked language model pre-trained on data in 100 different languages, making it highly effective for cross-lingual transfer learning. We initially experimented with both the Base (278M parameters) and Large (559M parameters) variants using Low-Ranking Adaptation (LoRA). Contrary to expectations, the Large model yielded almost no performance improvement over the Base model in our preliminary tests. However, we used XLM-R Large for the main architecture to be more competitive in the challenge.

The critical concerned about the use of LoRA was whether restricting updates to a small subset of parameters would limit the model's capacity to learn complex polarization patterns. To test this, we conducted a comparative experiment using full fine-tuning (updating 100% of the parameters). The results showed similar outcomes between the fully fine-tuned model and the LoRA-adapted model, confirming that updating only $\sim 2\%$ of the parameters is sufficient for this task, while significantly

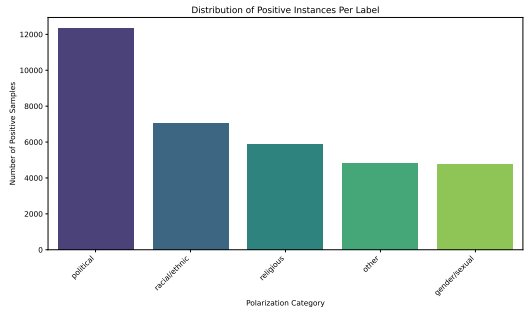

Figure 2: Distribution of polarization categories in Subtask 2.

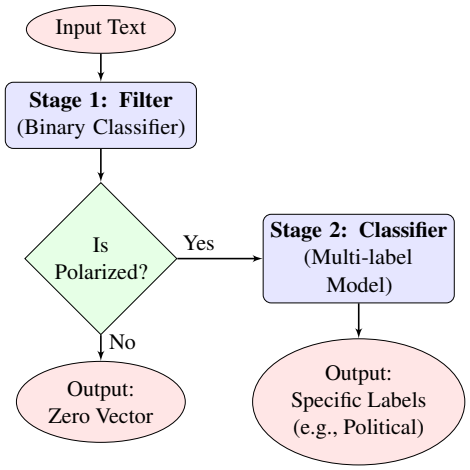

Figure 3: The Two-Stage Cascaded Pipeline architecture. Non-polarized texts are filtered out at Stage 1 to prevent false label attribution in Subtask 2.

reducing memory overhead.

4.1.1 Subtask 1: Weighted Binary Classification

For the binary detection task, we attached a standard classification head to the XLM-R encoder. While the dataset is globally balanced, specific languages exhibit significant skew. To mitigate this, we employed a Weighted Cross-Entropy loss. We computed class weights inversely proportional to class frequencies.

4.1.2 Subtask 2: Two-Stage Cascaded Pipeline

Initial experiments with a single multi-label model revealed a critical flaw: the model frequently predicted polarization types for texts that were non-polarized. To solve this, we implemented a Two-Stage Cascaded Pipeline:

Stage 1 (Filter): A binary classifier (identical to the Subtask 1 model) determines if the text contains polarized content.

Stage 2 (Classifier): if Stage 1 prediction is positive, the text passed to the second model, which predicts the specific category. Otherwise, the system outputs a zero vector.

This architecture ensures that the Stage 2 model focuses its capacity solely on distinguishing between polarization types without being confused by neutral content. The Stage 2 model was trained using Binary Cross-Entropy (BCE) with positive weights to address the scarcity of some labels:

$$\mathcal{L} = -\sum_{i=1}^{5} [w_i \cdot y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

4.2 Translation-based Cross-lingual Approach

To handle the multilingual nature of the dataset, we adopted a translation-based cross-lingual strategy focusing on a subset of 12 languages: English plus 11 languages translated into English using MarianMT. This setting corresponds to a translate-train pipeline, where all training and evaluation texts are mapped into a single target language. Prior 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 English dataset, enabling the use of a strong English pretrained language model, without relying on multilingual representations. We fine-tuned the model on the merged English dataset obtained after translation and we report results both on the aggregated dataset and through language-wise evaluation to quantify how translation affects performance across the 11 translated languages. While automatic translation may introduce noise or alter culturally specific expressions, most task-relevant semantic information is typically preserved, making translation a robust and effective strategy in cross-lingual classification under resource and modeling constraints.

4.3 Prompting-based strategy

As an alternative to the approaches presented so far, we explored a strategy based on the use of a non fine-tuned encoder-only model. Specifically, we selected LLaMA 3.1 8B Instruct, an instruction-tuned model designed to generate textual sequences while closely fitting to user-provided instructions.

To effectively leverage the model's generative capabilities, we introduced an explicit reasoning step. The model is required to provide a brief explana-

tion of why a given comment should be considered polarized or non-polarized, and subsequently output a set of boolean flags representing the final classification outcome.

In order to obtain coherent, interpretable, and easily processable outputs, we defined a structured JSON-style output format. The interaction with the model follows a few-shot approach, supplying a set of illustrative examples that the model can use as guidance. This choice proved to be crucial: a preliminary analysis showed that, in the absence of examples, the model struggled to consistently produce outputs that conformed to the required structure. Similarly, 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.

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.

The proposed architecture incorporates the insights discussed in the previous sections: comments are analyzed after being translated into English, and the binary polarization classification is treated as a prerequisite for category identification. To ensure consistency across predictions, the model output is post-processed by setting all category labels to False whenever a comment is classified as non-polarized. This design choice is motivated by the empirical observation that the model may occasionally assign a category to non-polarized comments, particularly when a specific topic is strongly represented in the text.

5 Experimental results

All tables in this section show the F1-score metric calculated as the arithmetic mean between languages.

5.1 Subtask 1 results

Table 2 shows the results of all three architectures proposed for subtask 1.

Architecture 1 Our experiments, on the XLM-R Base, revealed that increasing the model capacity via LoRA parameters yielded diminishing returns: increasing the LoRA rank from 16 to 32 resulted in no performance gain (with $F1_{macro}$ varying negligibly from 0.7805 to 0.7702). This

Example of Generative 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 4: 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.

Arch	F1-macro
XLM-RoBERTa Large	0.7930
MarianMT + deBerta v3	0.7060
LLaMA Instruct + Reasoning	0.6555

Table 2: Results of subtask 1

suggests that the lower rank is sufficient for the binary detection task. Furthermore, while the *XLM-R-Large* model achieved the highest absolute scores ($F1_{macro} = 0.7934$), the increment over the Base model was marginal. Consequently, the Base model with rank 16 remains the most efficient solution, offering competitive performance with significantly lower computational requirements.

Architecture 2 For the translation-based model, we conducted a limited hyperparameters tuning study by varying the learning rate and max length, but 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 ita ($F1_{macro} = 0.5983$) and urd ($F1_{macro} = 0.6011$) exhibited the largest drops. This dispersion is plau-

sibly 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.

Arch	F1-macro
Cascade PL	0.5630
MarianMT + deBerta v3	0.4842
LLaMA Instruct + Reasoning	0.6335

Table 3: Results of subtask 2

Architecture 1 For the single-model approach, we conducted a series of optimization experiments to improve performance through hyperparameter tuning. We tested configurations involving increased dropout, optimized class thresholds, and targeting both Attention and FFN layers for LoRA adaptation. Despite these extensive optimizations aimed at boosting the XLM-R Large model, we observed that these changes did not yield a substantial increment in performance. The $F1_{macro}$ score improved only slightly from 0.6798 to 0.7100, indicating a performance plateau where parameter tuning fails to provide significant gains. In contrast, the Cascade strategy demonstrated a distinct behavior: while the Accuracy was lower than the optimized single model, the pipeline achieved a significantly higher $F1_{macro}$ (0.7918), suggesting that a serialized architecture is more effective at capturing minority labels than a heavily tuned single large model.

Architecture 2 For the translation-based approach in Subtask 2, 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. This variability highlights the non-uniform effects of translation across languages, especially in a multi-label setting where subtle semantic shifts may impact category-specific predictions.

5.3 Architecture 3 results

We explored the LLaMA 3.1–based approach by two distinct experiments, which share the same prompt structure and the same set of examples, differing only in the presence of a rationale field in the output.

In Task 1, the model is required to generate, in addition to the classification labels, a short textual explanation (rationale) motivating the polarization decision.

In Task 2, the prompt does not include the rationale field, and the model outputs only the boolean flags related to polarization and its categories. Apart from this difference, the prompt content, output format, and contextual instructions remain identical to the first tasks.

	Classification only	Reasoning and Classification
F1-macro (task 1)	0.5999	0.6555
F1-macro (task 2)	0.5963	0.6335

Table 4: Comparison of Architecture 3 performance with and without Reasoning

The experiments we conducted allowed us to assess the effectiveness of introducing a reasoning step compared to direct classification. Using the same test set (constructed by selecting 10% of the comments for each Language) we observed that simply adding the rationale generation in Task 1 leads to slightly better overall results than Task 2. This suggests that a richer response, in which the model explicitly articulates its reasoning process, can positively influence the subsequent assignment of classification labels.

It is interesting noting that, although rarely, the model may produce structurally invalid outputs. For instance, it may refuse to respond when faced with 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 and restrictive instructions within 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 Conclusion and future works

What emerges clearly from our results is the different effectiveness of the strategies adopted in the two tasks considered. Specifically, the first architecture achieves the most promising solution for binary polarization detection, whereas the third architecture achieves superior results in the categorization task. On the other hand, the second architecture consistently underperforms in both scenarios; this suggests that the translation process acts as an information bottleneck, eroding the specific stylistic and semantic markers necessary to identify polarized speech.

The performance dichotomy highlights a significant trade-off between architectural efficiency and semantic reasoning capabilities. For the binary detection task (Subtask 1), the XLM-RoBERTa approach proved to be the most robust. Its success suggests that an optimized multilingual encoder remains an optimal solution to efficiently identify the presence of polarized content. The model prevents false positives without the need for the heavy computational overhead associated with generative models.

Conversely, the superior performance of the Generative LLaMA-based approach in the categorization task (Subtask 2) indicates that detecting the type of polarization requires a deep semantic understanding that goes beyond simple pattern matching. The integration of an explicit reasoning phase allowed the model to leverage broader contextual knowledge, articulating the specific nature of hostility more effectively than the encoder. This confirms that decoders are particularly well-suited for tasks where the decision boundary is based on subtle contextual or cultural nuances.

Finally, the limitations observed in the translation-based approach reinforce the hypothesis that polarization is deeply rooted in the original linguistic and cultural contexts. The translation process seems to introduce noise and flatten specific idiomatic expressions, reducing the model's ability to detect hostility.

An interesting direction for future work involves the adoption of a collaborative approach, in which the first architecture is responsible for binary polarization classification, while the second architecture is employed for assigning the specific polarization categories. This strategy would make it possible to use each model where it performs best.

An alternative option would be to focus on a single model by combining the strategies explored in the three architectures. A promising starting point in this direction is the approach proposed in (Wenna Lai, 2025), where an encoder–decoder model is fine-tuned on a dataset augmented with artificially generated rationales. Starting from the available dataset, rationales could be generated through a rationalization technique, asking a generative model to explain why a comment is considered polarized or not, while simultaneously providing the ground-truth label. The resulting data could then be used to train a new model, with the goal of internalizing not only classification capabilities but also task-specific reasoning related to polarization.

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.

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*.

Saba Anwar Sarah Kohail Rudy Alexandro Garrido Veliz Robert Geislinger Aisha Jabr Idris Abdulmumin Laiba Qureshi Aarushi Ajay Borkar Maryam Ibrahim Mukhtar Abinew Ali Ayele Ibrahim Said Ahmad Adem Ali Martin Semmann Shamsuddeen Hassan Muhammad Seid Muhie Yimam Usman Naseem, Juan Ren. 2026. Task 9 of polar @ semeval: Detecting multilingual, multicultural, and multievent online polarization. arXiv preprint arXiv:2505.20624 (Updated Jan 2026).

Guandong Xu Qing Li Wenna Lai, Haoran Xie. 2025. Rvisa: Reasoning and verification for implicit sentiment analysis. *IEEE Transactions on Affective Computing*.

A Appendix A: Per language results

Language	XLM-RoBERTa Large	MarianMT + deBERTa v3 Base	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 5: F1-macro per language for Subtask 1.

Language	XLM-RoBERTa Large	MarianMT + deBERTa v3 Base	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 6: F1-macro per language for Subtask 2.