pfr821 at SemEval-2026 Task 9: Multilingual Polarization Detection via Hybrid XLM-RoBERTa with Targeted Data Augmentation and Imbalance-Aware Training

Antoine Durand[†], **Rémi Hamon**[†], **Matthieu Pereira**[†], **Nathan Boucneau**[†], **Paul Cintra**[†], **Guillaume Gadek**[⋆], **Louis Lefebvre**[⋆], **Matthieu Labeau**[†]

[†]Télécom Paris, Institut Polytechnique de Paris [⋆]Airbus Defence & Space

durantoine.d@gmail.com, remi.hamon2022@gmail.com, matthieu.pereira@telecom-paris.fr, cintrapaul@gmail.com

nathan.boucneau@gmail.com, guillaume.gadek@airbus.com, louis.lefebvre@airbus.com, matthieu.labeau@telecom-paris.fr

Abstract

This paper describes **HYPOLDET**, the system submitted by team **pfr821** to SemEval-2026 Task 9 (Polarization Detection, Subtask 1), a binary classification task over 22 typologically diverse languages. Our approach combines three complementary contributions. We first extend XLM-RoBERTa-Large with a custom transformer encoder layer and a learned attention-based pooling mechanism (*Hybrid Architecture*), allowing the model to aggregate token-level signals beyond the `[CLS]` representation. We then augment training data through a targeted LLM-based synthetic generation pipeline (Grok API), producing culturally grounded examples for low-resource and imbalanced languages. Finally, independently of data augmentation, we address class imbalance at the training level through an imbalance-aware regime combining a per-language balanced batch sampler, weighted focal loss, and label smoothing. Our best single model achieves an unweighted macro-averaged F1 of **0.796**, and a lightweight ensemble reaches **0.798**, ranking in the top 10 for 7 languages and 2nd place for Hausa.

1 Introduction

Polarization in online discourse, understood as the amplification of divisive, extremist, or ideologically charged language, poses significant risks to democratic processes and social cohesion (Garimella et al., 2018). Addressing it automatically across diverse languages remains a core NLP challenge. SemEval-2026 Task 9 (Naseem et al., 2026b), based on the POLAR benchmark (Naseem et al., 2026a), provides a large-scale multilingual testbed for this problem, covering 22 languages spanning several scripts and resource levels.

Subtask 1 is formulated as binary classification: given a text fragment, predict whether it is *polarized* (1) or *non-polarized* (0). Performance is measured by per-language macro F1 averaged over both classes, and the macro-average over all 22 languages serves as the global summary metric, giving equal weight to every language regardless of data volume.

The POLAR dataset presents three compounding challenges: **class imbalance**, **linguistic heterogeneity** across diverse scripts, morphologies, and discourse norms, and **data scarcity**, with some languages providing fewer than 1,700 training examples.

A core design principle of **HYPOLDET** is to produce a **single, language-agnostic model** covering all 22 languages simultaneously, with no per-language components. This global approach is motivated by the cross-lingual transfer properties of multilingual pre-trained models: shared representations across related languages and scripts reduce the per-language data requirement, while the added task-specific components (Section 2.1) remain lightweight relative to the backbone.

HYPOLDET combines three complementary components: a **Hybrid XLM-RoBERTa-Large** architecture with a task-specific encoder layer and attention-based pooling; an **LLM-based synthetic data pipeline** targeting low-resource and imbalanced languages; and an **imbalance-aware training regime**.

We build on XLM-RoBERTa-Large (Conneau et al., 2020), whose cross-lingual representations have proven effective across typologically diverse languages. Our approach draws on LLM-based synthetic data generation (Gilardi et al., 2023) for low-resource augmentation, and Focal Loss (Lin et al., 2017) for imbalance-aware training. Code and models are publicly available at `https://github.com/Projet-Fil-Rouge-Airbus-Propagande/online-polarization-detector`.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1169–1174

2 System Overview

2.1 Architecture: Hybrid XLM-RoBERTa-Large

Our model, **HybridXLM-RoBERTa**, extends XLM-RoBERTa-Large (`xlm-roberta-large`: 24 layers, 1024 hidden size, 16 attention heads, fine-tuned end-to-end) with two trainable components stacked on top of the base encoder: a task-specific transformer block and an attention-based pooling mechanism.

Additional Transformer Encoder Layer. A single custom `TransformerBlock` is applied on top of the XLM-RoBERTa-Large output sequence, containing two sub-layers: a **multi-head self-attention** module with the number of heads selected from $\{2, 4, 8\}$ by Optuna (hidden size 1024), and a **feed-forward network** with intermediate size 512 and ReLU activation. Each sub-layer uses Post-LN residual connections. This block adds approximately 5.3M parameters (≈1.5% of the backbone), allowing the model to learn task-specific interaction patterns without disrupting the multilingual representations acquired during pretraining.

Attention-Based Pooling. Rather than relying on the `[CLS]` token or mean pooling, we apply a learned attention mechanism that produces a weighted sum of the full token sequence:

$$e_i = \mathbf{w}^\top \tanh(\mathbf{W}\mathbf{h}_i + \mathbf{b}) \quad (1)$$

$$\alpha_i = \mathrm{softmax}(e_i) \quad (2)$$

$$\mathbf{s} = \sum_i \alpha_i \cdot \mathbf{h}_i \quad (3)$$

where $\mathbf{h}_i \in \mathbb{R}^{1024}$ are the token representations from the extra encoder layer. This allows the model to selectively focus on the most informative tokens for polarization, regardless of their position in the sequence.

Projection Head. The 1024-dimensional pooled representation is further transformed through $\mathrm{Linear}(1024 \to 256) \to \mathrm{LayerNorm} \to \mathrm{GELU} \to \mathrm{Dropout}$ before the final binary classifier. Figure 1 provides an overview of the complete architecture.

2.2 Synthetic Data Generation

Several languages in the dataset exhibit severe class imbalance, reducing the effective training signal for the minority class. To address this, we build a targeted synthetic data generation pipeline using the

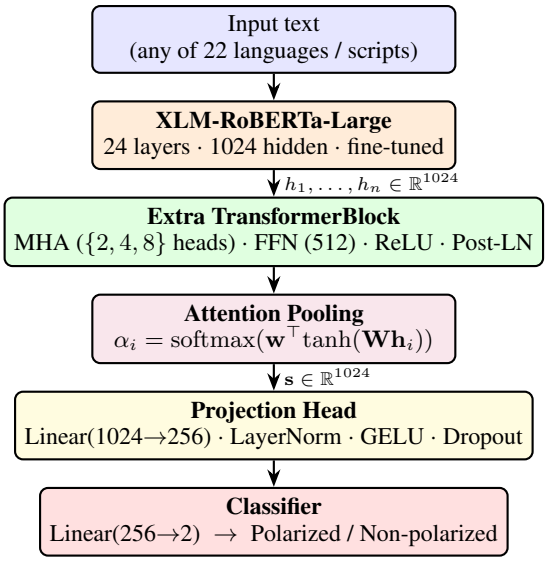

Figure 1: Architecture of HybridXLM-RoBERTa. A single model is trained across all 22 languages simultaneously, with no language-specific components.

Grok API (xAI). This API was used exclusively for academic research within the SemEval-2026 shared task framework.

Generation Protocol. Each request is conditioned on multiple axes to maximise diversity:

- **Content**: 6 polarization categories (political, ethnic, religious, gender, etc.) with 10 subtopics each; 49 neutral topics for the non-polarized class to avoid topical artifacts;
- **Style and intensity**: 14 tweet styles crossed with three intensity levels (light sarcasm, moderate accusations, strong hostile rhetoric);
- **Language realism**: script conventions, code-switching, language-specific politeness forms, and punctuation cues for all 22 languages; 10–20% noise (typos); realistic length distribution; sampling temperature in $[0.75, 1.05]$; 4 prompt variants rotated across batches.

Training Integration. At training time, the proportion of synthetic to real examples per batch is controlled by a `synthetic_ratio` hyperparameter tuned by Optuna. The synthetic subset is resampled with a different random seed each epoch (`DynamicSyntheticDataset`), maximising exposure to the full diversity of generated examples. Languages with severe class imbalance or low training volume receive larger synthetic volumes. In total, 21,223 synthetic examples were generated across all 22 languages, with the highest volumes

for Khmer (4,825) and Hausa (2,279). Table 1 illustrates representative generated examples.

Label	Generated example (English)
polar.	*Why is it that every protest turns into a riot when the right shows up, but lefties get a free pass? Double standards much?*
polar.	*Leftie hypocrites preaching climate doom from their private jets. Elites gonna elite, right? #Hypocrisy*
non-p.	*Cup of tea with custard creams — best afternoon treat.*

Table 1: Illustrative LLM-generated synthetic examples. Polarized examples span political and ideological types; non-polarized examples are topically neutral.

2.3 Imbalance-Aware Training

Beyond data augmentation, we address class imbalance at the training level through three complementary mechanisms.

Per-Language Balanced Batch Sampler. We implement a **per-language balanced sampler** that ensures equal class representation within each mini-batch (Buda et al., 2018). For each language l, sampling weights are set inversely proportional to class frequency, and samples are drawn with replacement. This prevents the gradient signal from being dominated by the majority class in high-resource, imbalanced languages.

Focal Loss. We additionally explore **per-language focal loss**:

$$\mathcal{L} = -(1 - p_t)^{\gamma} \cdot \log(p_t) \cdot w_{l,c} \tag{4}$$

where p_t is the predicted probability for the true class and γ controls focusing strength. The **class-weight** term $w_{l,c}$ is a language- and class-specific weight derived from training-set frequencies, which handles class imbalance reweighting directly within the loss.

Label Smoothing. To reduce overconfidence, we apply label smoothing with factor ϵ tuned via Optuna, which acts as a regularizer particularly relevant given potential label noise in synthetic data.

2.4 Hyperparameter Optimization

Given the large number of interacting hyperparameters, we rely on **Optuna** (Akiba et al., 2019) for automated hyperparameter search, with each trial training a full model and reporting validation macro F1. The search space covers:

- **Learning rate** $[4.0\times10^{-6},\ 5.5\times10^{-6}]$ (log scale): XLM-RoBERTa-Large is sensitive to catastrophic forgetting at higher rates;
- **Weight decay** $[0.15,\ 0.40]$: regularizes the large parameter count of the backbone;
- **Training epochs** $\{8, \ldots, 14\}$: balances convergence with early stopping (patience 3);
- **Label smoothing** $\epsilon \in [0.16,\ 0.30]$: calibrates confidence against noisy synthetic labels;
- **Dropout** (base / extra layer) $[0.18, 0.40]$ / $[0.20, 0.40]$: jointly controls regularization;
- **Synthetic ratio** $[0.20,\ 1.0]$: governs the proportion of synthetic examples per training batch;
- **Focal loss** $\gamma \in \{0.0,\ 2.0,\ 2.5,\ 3.0\}$: explores the spectrum from standard CE to aggressive focusing.

Trials are persisted in an SQLite database and tracked with MLflow. Due to compute constraints, the search was not exhaustive; all results reflect the best configuration found within the available budget. The optimizer is AdamW with linear warm-up over 8% of training steps, followed by linear decay and gradient clipping at norm 1.0.

2.5 Ensemble and Model Selection

Our primary objective was to produce a single well-optimized model. Ensembling was introduced late, once it became clear that the Optuna search budget would not suffice before the submission deadline. Final class probabilities are obtained by **uniform probability averaging** over the top-N configurations ranked by validation macro F1. The gain is marginal (≈**+0.002 macro F1**), as shared backbone and data yield highly correlated predictions.

3 Experimental Setup

3.1 Data

The official SemEval-2026 Task 9 Subtask 1 training set comprises **80,377 samples** across 22 languages. Class distributions vary substantially: some languages (Arabic, Chinese, Swahili) are relatively balanced, while others (Khmer, Hausa, Hindi) exhibit severe imbalance with the polarized class representing fewer than 30% of samples. We

supplement the official training data with LLM-generated synthetic examples as described in Section 2.2, and use a stratified **20%** hold-out split of the *real* training data as an internal validation set for early stopping and Optuna objective evaluation; synthetic examples are never included in this split.

3.2 Training Infrastructure

Models are trained on two NVIDIA P100 GPUs using Distributed Data Parallel (DDP) via PyTorch `torchrun`, with gradient accumulation of 2 steps for an effective batch size of 16 (8 per device). Optuna trials ran sequentially, each leveraging both GPUs. All experiments are tracked in MLflow, logging per-epoch metrics and per-language F1 scores.

4 Results

4.1 Official Competition Results

Our system was evaluated on the official SemEval-2026 Task 9 Subtask 1 test set; per-language F1 scores and rankings are reported in Table 2.

Our ensemble achieves a macro-averaged F1 of **0.798** (best single model: **0.796**) across all 22 languages, placing in the **top 10 for 7 languages** and finishing **2nd for Hausa**. F1 scores range from 0.901 for Nepali to 0.633 for Italian, reflecting considerable variation in task difficulty across languages.

4.2 Analysis

Languages with class imbalance. Our system performs particularly strongly on initially imbalanced languages: Hausa (2nd), Odia (7th), and Amharic (5th). The balanced sampler, class-weighted focal loss, and synthetic augmentation all target this challenge and act in the same direction, making it difficult to attribute results to any single mechanism. That said, Hausa benefits from one of the highest synthetic augmentation ratios and achieves our best relative ranking, which is consistent with augmentation intensity playing a role, though we cannot isolate this effect from the other mechanisms.

Error analysis. *Content note:* the following analysis reproduces verbatim examples from the POLAR dataset that readers may find offensive, included solely to characterize error patterns; we do not endorse any of the views expressed.

Italian (0.633), German (0.732), and Spanish (0.763) all fall below the global macro-average (0.798), each with a distinct failure signature.

Italian is the most problematic language: the model classifies only 29.6% of examples as polarized against an actual rate of 47.3%. The failure is driven by implicit framing carrying no surface hate marker, as in *"7 minuti di servizio e non sono riusciti a dire ROM, complimenti"* ("Seven minutes of coverage and they couldn't manage to say 'Roma', well done"), a polarized post predicted as non-polarized where hostility is encoded entirely in pragmatic implication: the claim that ethnicity was the central omitted fact, with no hostile lexical item present.

Spanish shows the opposite asymmetry (214 FP vs. 138 FN): identity tokens (*judio*, *lesbiana*, *musulman*) trigger the classifier regardless of context, misflagging reclaimed slang (*"que bien tengo el pelo maricon"*: "my hair looks so good, queer") as polarized, while culture-specific signals such as regional slurs (*"jopo jopo"*, a derogatory indigenous term) are missed entirely.

German has a more balanced profile (182 FP, 200 FN). The most striking false negatives involve historical dog-whistles with no surface marker, such as *"Wie ist denn der Überbegriff von Juden, Schwulen, Behinderten und Kommunisten?"* ("What is the umbrella term for Jews, gays, disabled people, and communists?"), which frames an enumeration of Nazi victim categories as an innocent quiz.

Across all three languages, the model conflates *topic mention* with *polarization stance*, over-triggering on sensitive lexical items while missing hostility expressed through framing, implication, or cultural convention. These patterns are consistent with a structural hypothesis: a single shared representation space may struggle to capture the rhetorical and cultural specificity of each language's political discourse.

4.3 Ablation Study

Following reviewer feedback, we conducted an ablation study to assess each component's contribution. Each variant is independently optimized with Optuna for 20 trials with a rotating train/validation seed; results should be treated as indicative given the limited budget.

Architecture. We compare a **baseline** (XLM-RoBERTa-Large with `[CLS]` pooling) against two incremental extensions: adding attention-based pooling and a projection head, then further adding an extra TransformerBlock (full hybrid). Each step

Language	F1	Rank	/Total	Language	F1	Rank	/Total
Hausa (hau)	0.832	**2**	/31	Russian (rus)	0.786	13	/31
Amharic (amh)	0.786	5	/30	Burmese (mya)	0.873	18	/30
German (deu)	0.732	5	/33	Persian (fas)	0.802	19	/32
Odia (ori)	0.805	7	/33	Khmer (khm)	0.709	19	/31
Turkish (tur)	0.797	7	/31	Chinese (zho)	0.891	19	/33
Arabic (arb)	0.832	8	/33	Bengali (ben)	0.829	21	/37
Polish (pol)	0.811	8	/32	Hindi (hin)	0.793	22	/35
Telugu (tel)	0.883	12	/33	Spanish (spa)	0.763	23	/37
Urdu (urd)	0.787	11	/35	Punjabi (pan)	0.753	25	/33
Italian (ita)	0.633	12	/32	English (eng)	0.785	29	/45
Nepali (nep)	**0.901**	16	/33	Swahili (swa)	0.768	26	/31
Macro-avg F1 (22 languages): 0.798 (ensemble) / 0.796 (best single model)							

Table 2: Per-language results for team pfr821 on the SemEval-2026 Task 9 Subtask 1 test set (official rankings). Bold: best rank / best absolute F1.

yields a slight improvement in best-trial performance ($+0.24\%$ per step, $+0.50\%$ total on test). A more striking effect is the stabilization of training brought by attention-based pooling: test-set trial variance drops from 0.017 to 0.005 and remains low throughout the full hybrid (0.005), suggesting that the architecture primarily reduces poorly-converging trials. This stabilization is achieved with a parameter overhead of only 5.3M ($\approx$1.5% of the backbone).

Synthetic data. Adding a **full hybrid without synthetic data** condition reveals a distribution effect: without synthetic examples, training and validation share the same distribution, which mechanically inflates validation F1 in that condition. On the test set, best-trial performance is comparable with and without synthetic data, but the val-test gap is narrower with synthetic examples, consistent with the hypothesis that they provide targeted coverage of imbalanced languages while introducing some distributional shift. We also note that the 20-trial Optuna budget had not fully explored the synthetic ratio in this condition, so the ablation likely underestimates the contribution of synthetic augmentation.

5 Conclusion

We presented **HYPOLDET**, our system for SemEval-2026 Task 9 Subtask 1, built on three complementary contributions: a Hybrid XLM-RoBERTa-Large architecture with a task-specific encoder layer and attention-based pooling, a targeted LLM-based synthetic data pipeline, and an imbalance-aware training regime. The best single model achieves an unweighted macro-averaged F1 of **0.796**, while a lightweight ensemble reaches

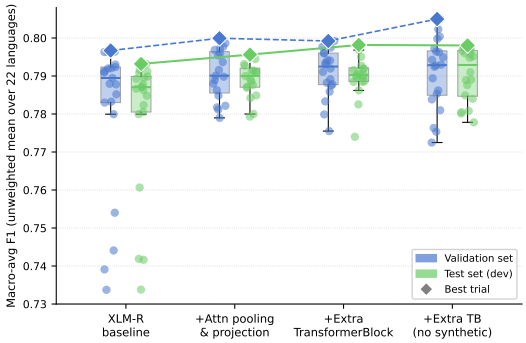

Figure 2: Incremental architecture and synthetic data ablation. Unweighted macro-averaged F1 across Optuna trials (rotating train/val seeds); diamonds mark the best trial per variant. Solid line: test set (dev); dashed line: validation set.

0.798, ranking in the top 10 for 7 of 22 languages and 2nd for Hausa.

The ablation suggests that no single contribution dominates: each mechanism makes a marginal but consistent contribution. The role of synthetic data in particular warrants further investigation: whether it genuinely improves coverage for imbalanced languages or primarily introduces distributional noise that trades off against overall quality remains an open question.

Several directions remain open. Better quality filtering of synthetic examples would reduce distributional shift from generated data. Finally, domain-adaptive pre-training of the task-specific layers on unlabelled social media corpora, prior to fine-tuning, could specialise them on the informal and polarization-prone register of online discourse without disturbing the backbone's cross-lingual representations.

Ethical Considerations

Use of the Grok API. Synthetic training data was generated via the Grok API (xAI) exclusively for participation in the SemEval-2026 shared task. The generated content was used solely as training material and is not intended for redistribution or deployment outside this academic context.

Offensive content in examples. This paper reproduces verbatim examples drawn from the POLAR benchmark dataset to illustrate error patterns in our analysis. Some of this content may be perceived as offensive. We include these examples solely to better characterize the difficulty of the task and do not endorse any of the views expressed.

Dual-use risk. A polarization classifier could potentially be misused to suppress legitimate political speech or to target specific communities. Automatic classifiers should not be used as sole arbiters in content moderation decisions; human oversight remains essential.

Acknowledgments

This work was carried out as part of the *Mastère Spécialisé* in Artificial Intelligence at Télécom Paris, Institut Polytechnique de Paris. The project originates from a collaboration with Airbus Defence & Space, whose industrial framing shaped both the problem definition and its applicative scope.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of KDD 2019*.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1).

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV 2017*.

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, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026a. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

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, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. 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.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of SemEval-2023*.