Semantic Vectors at SemEval-2026 Task 9: Robust Multilingual Polarization Detection via Dual-Encoder Fusion and Expert Ensembling

Ankit Dash* Priyanshu Mittal† Piyush Prashant† Sunil Saumya†
*Department of Computer Science & Engineering
†Department of Data Science and Artificial Intelligence
Indian Institute of Information Technology Dharwad, Karnataka, India
24bcs016@iiitdwd.ac.in, 24bds058@iiitdwd.ac.in,
24bds055@iiitdwd.ac.in, sunil.saumya@iiitdwd.ac.in

Abstract

We present SEMANTICVECTORS, our system for POLAR@SemEval-2026 Task 9 on multilingual online polarization detection across 22 typologically diverse languages. Polarization is frequently conveyed through implicit rhetorical framing, making cross-lingual detection highly challenging. We address this with a *Siamese dual-encoder* jointly fine-tuning mDeBERTa-v3-base and XLM-RoBERTa-large via 4-bit QLoRA, fused with language-specific expert models (GBERT, Italian BERT, Swahili BERT) through an XGBoost meta-stacker with per-language Platt calibration. Rather than addressing class imbalance, focal loss functions as a *hard-example miner*, concentrating gradients on subtly framed instances rather than lexically obvious ones. Combined with per-language threshold optimization, our system achieves macro-F1 = **0.797** and accuracy = **0.827** across all 22 languages.

1 Introduction

Online polarization—defined as the sharp, hostile division between social, political, or identity groups in digital discourse—has emerged as a major driver of societal fragmentation worldwide (Tucker et al., 2018; Naseem et al., 2026b). Unlike hate speech, which typically relies on explicit slurs or direct incitement, polarization manifests through subtler mechanisms: the stereotyping and vilification of out-groups, blind in-group solidarity, and the gradual normalization of dehumanizing language (Naseem et al., 2026b). These properties make it a distinct and under-studied NLP challenge—polarized content often *passes* standard toxicity filters, yet systematically erodes social cohesion when left undetected at scale.

The computational difficulty is compounded in multilingual settings. Divisive framing is deeply culture-specific: a rhetorical question signalling contempt in German political discourse may appear syntactically neutral in isolation; an ironic phrase marking in-group polarization in Hindi carries no such signal in direct translation. Prior benchmarks have largely focused on English or a small set of high-resource languages, leaving cross-lingual polarization detection without a rigorous, large-scale evaluation framework.

POLAR@SemEval-2026 Task 9 addresses this gap with the first shared benchmark for *multilingual, multicultural, and multievent* polarization across 22 typologically diverse languages spanning six language families. We participate in Subtask 1 (binary: polarized vs. neutral; metric: macro-averaged F1), which requires models that generalize across radically different scripts, morphological structures, and cultural framings of divisiveness.

Our system, SEMANTICVECTORS[1], tackles these challenges through three coordinated components: (1) a *Siamese dual-encoder* jointly fine-tuning mDeBERTa-v3-base (He et al., 2021) and XLM-RoBERTa-large (Conneau et al., 2020) via 4-bit QLoRA (Dettmers et al., 2023), capturing complementary syntactic and semantic inductive biases; (2) language-specific expert models for German, Italian and Swahili, fused through an XGBoost meta-stacker (Chen and Guestrin, 2016) with Platt calibration (Platt et al., 1999); and (3) per-language threshold optimization on development data.

Crucially, we highlight two novel contributions to multilingual affective computing: to our knowledge, this is the first architecture to explicitly fuse disjoint subword tokenization spaces (SentencePiece and BPE) as a structural solution to cross-lingual out-of-vocabulary bottlenecks, and the first to deploy Shannon entropy as a dynamic gating signal for routing text to cultural expert models.

2 Background and Related Work

Task and Data. POLAR@SemEval-2026 Task 9 (Naseem et al., 2026a,b) provides per-language training, development, and test splits. The pooled training corpus contains ≈40,395 samples across 22 languages with mild class imbalance (class

[1]Code: https://github.com/AnkitDash-code/Semantic-Vectors-SemEval

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1903–1910

Family	Languages
Western European	deu, eng, ita, spa, pol
Indic / Indo-Aryan	hin, nep, ori, pan, tel, urd, ben
Semitic	arb, fas
African	amh, hau, swa
Southeast Asian	khm, mya
Sinitic / Turkic / Slavic	zho / tur / rus

Table 1: The 22 target languages grouped by typological family.

weights: neutral = 1.042, polarized = 0.961). Languages span six typological families (Table 1). Resource levels vary enormously: English provides thousands of training examples while Hausa and Odia have fewer than 500, creating non-uniform generalization constraints across the language set.

Related Work. Polarization detection is informed by stance detection (Mohammad et al., 2016) and computational framing analysis (Card et al., 2015). Our encoders belong to the Transformer family (Vaswani et al., 2017), with BERT (Devlin et al., 2019) as the seminal pre-trained backbone. XLM-RoBERTa (Conneau et al., 2020) is the standard multilingual backbone; mDeBERTa-v3 (He et al., 2021) extends disentangled attention. Siamese networks (Bromley et al., 1993; Koch et al., 2015) leverage shared-weight twin encoders for dense representations. QLoRA (Dettmers et al., 2023) and LoRA (Hu et al., 2022) enable large-encoder fine-tuning under hardware constraints; focal loss (Lin et al., 2017) re-weights training examples by prediction difficulty, enabling focus on hard-to-classify instances.

3 System Architecture

Our architecture (Figure 1) comprises five stages: (1) Siamese dual-encoder training; (2) Focal Loss with Label Smoothing; (3) language-specific expert Routing; and (4) XGBoost meta-stacking with Platt calibration; and (5) Per-Language Threshold Optimization

3.1 Siamese Dual-Encoder

mDeBERTa-v3-base (183M params) (He et al., 2021) uses ELECTRA-style pre-training with gradient-disentangled attention, separating content from positional representations, thereby enabling finer-grained token-level disambiguation. We apply LoRA adapters (r=32, α=64) to `query_proj` and `value_proj`. This model is incompatible with Flash Attention 2; we use PyTorch SDPA and `bfloat16` precision (native H100, more stable than `float16`).

XLM-RoBERTa-large (560M params) (Conneau et al., 2020) provides broad cross-lingual coverage over 100 languages. Flash Attention 2 (Dao, 2023) is used when available (SDPA fallback). LoRA (r=32, α=64) on `query`/`value`. Both encoders process the input independently; their `[CLS]` representations are concatenated:

$$\mathbf{h} = [\mathbf{h}_A;\, \mathbf{h}_B] \in \mathbb{R}^{768+1024} = \mathbb{R}^{1792} \quad (1)$$

and passed through an MLP (LayerNorm (Ba et al., 2016) → Dropout(0.1) → GELU → Linear) to yield p_{hybrid}. Both encoders use 4-bit NF4 double quantization; gradient checkpointing is disabled due to instability with 4-bit quantized layers.

Fusing two distinct subword tokenizers also provides a practical vocabulary benefit: mDeBERTa's SentencePiece and XLM-R's BPE vocabulary are complementary, together covering a wider range of morphological decompositions. For complex, morphologically rich scripts such as Amharic (Ethiopic syllabary) and Odia (abugida), we hypothesize that a token unknown to one tokenizer's vocabulary is frequently handled by the other, reducing effective out-of-vocabulary (OOV) rate and improving subword coverage across low-resource African and Indic languages.

3.2 Focal Loss with Label Smoothing

$$\mathcal{L} = -\sum_i \sum_c (1 - p_{t,i})^\gamma \, \tilde{y}_{i,c} \log p_{i,c} \quad (2)$$

where $\tilde{y}$ is the label-smoothed target (ε=0.05) (Müller et al., 2019). The focusing parameter γ=2.0 acts as a *hard-example miner*: with an overall class imbalance ratio of only 1.08:1, the primary justification for focal loss is not label frequency but *instance difficulty*. Polarized content that uses implicit rhetorical framing (dog-whistles, sarcasm, rhetorical questions) yields high-entropy predictions under standard cross-entropy; γ=2.0 down-weights easy, explicitly polarized examples by up to 96%, forcing the model to allocate capacity to these hard, rhetorically framed instances (Lin et al., 2017).

3.3 Language-Specific Expert Routing

To counteract the dominant gradient updates from high-resource languages, we implement a selective expert routing mechanism. We selected German, Italian, and Swahili for dedicated expert routing based on distinct failure modes observed in our early dual-encoder baselines. German and Italian exhibited severe recall deficits (e.g., Italian baseline recall = 0.467) because their polarization frequently relies on implicit, ironic political framing that the global hybrid model failed to capture. Swahili was selected as a representative low-resource, morphologically complex language where the global backbone lacked the localized sociopolitical vocabulary to accurately parse divisive intent.

German texts are routed to a GBERT-base architecture Chan et al. (2020), Italian to an Italian-BERT

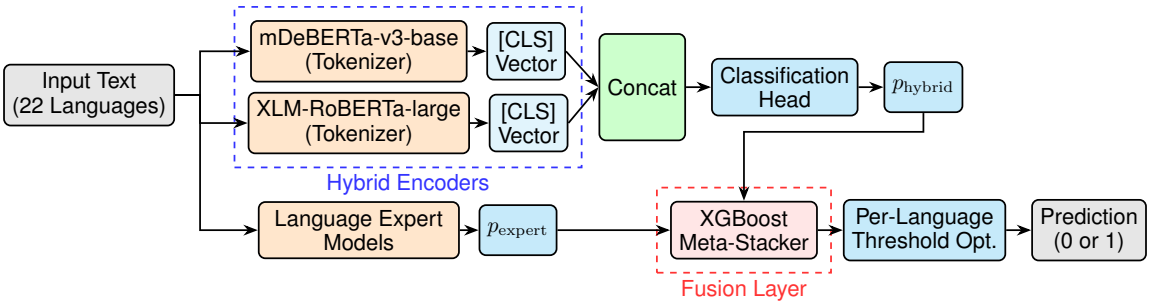


Figure 1: The proposed Multilingual Hybrid Architecture. mDeBERTa-v3-base and XLM-RoBERTa-large process input text in parallel; their `[CLS]` vectors are concatenated and passed through a Classification Head to yield p_{hybrid}. Language Expert Models produce $p_{\text{expert}}(\ell)$, fused via an **XGBoost Meta-Stacker**. Per-language threshold optimization yields $\hat{y} \in \{0, 1\}$.

Schweter (2020), and Swahili to a specialized Swahili-BERT(Adelani, 2022). Crucially, these expert models do not run universally across the dataset. During inference, they are dynamically triggered using the explicit language-ID tag provided in the task metadata. This conditional routing ensures that an expert model is only executed for texts matching its designated language subset, bypassing unnecessary computation for out-of-scope languages.

3.4 XGBoost Meta-Stacker with Platt Calibration

For targeted linguistic clusters where the Siamese backbone showed systematic underperformance, we train lighter QLoRA experts (r=16, α=32): GBERT-base (Chan et al., 2020) for German, which captures culturally specific ironic register; DBMDZ Italian BERT (Schweter, 2020) for Italian, which handles the implicit framing patterns responsible for *ita*'s low baseline recall (0.467); and a specialized Swahili-BERT(Adelani, 2022) to address critical localized sociopolitical vocabulary gaps in this morphologically complex, low-resource setting.

Predictions from the Siamese backbone and language expert are fused via an **XGBoost meta-stacker** (Chen and Guestrin, 2016) trained on out-of-fold (OOF) predictions following the stacked generalization paradigm (Wolpert, 1992). Rather than simple probability averaging, the stacker ingests a five-dimensional feature vector:

$$\mathbf{x}_{\text{meta}} = \left[p_{\text{hyb}},\ p_{\text{exp}},\ l_{\text{tok}},\ \max(p),\ H(p)\right] \quad (3)$$

where $p_{\text{hyb}}, p_{\text{exp}} \in [0, 1]$ are the predicted probabilities of the polarized class, l_{tok} is the token count, $\max(p)$ is the prediction confidence, and $H(p) = -\sum_c p_c \log p_c$ is Shannon entropy over class probabilities. Entropy acts as a reliability signal: high-entropy predictions indicate model uncertainty, prompting the stacker to weight the expert's output more heavily.

Following stacking, a per-language **Platt scaling** step (Platt et al., 1999) fits a logistic regression on XGBoost OOF outputs to calibrate posterior probabilities before thresholding. This two-stage design yields better-calibrated posteriors than raw averaging, particularly for low-resource languages with skewed class priors.

3.5 Per-Language Threshold Optimization

We grid-search 81 thresholds $t \in [0.1, 0.9]$ on the development set:

$$t^* = \arg\max_t \ \text{MacroF1}\big(\hat{Y}(t),\ Y_{\text{dev}}\big) \quad (4)$$

where $\hat{y} = \mathbf{1}[p_{\text{cal}} > t^*]$. Languages with $<$10 development examples default to t=0.5.

3.6 Negative Results and Architecture Search

To contextualize our final design, we summarize three explored alternatives that underperformed. **Linear-time architectures**: RWKV (Peng et al., 2023) and Mamba (Gu and Dao, 2023) achieved 2$\times$ training speedup but degraded on zero-shot cross-lingual transfer across 22 typologically diverse languages. **Extreme quantization**: BitNet 1.58-bit (Wang et al., 2023) attained macro-F1 = 0.977 on English in isolation, but generalized poorly multilingually due to representational collapse under ternary weights. **Data augmentation**: Easy Data Augmentation caused semantic drift in morphologically complex non-English languages. These results motivate choosing representational breadth over efficiency.

4 Experimental Setup

Data & Preprocessing. All official per-language training CSVs are pooled (≈40,395 samples). No external corpora, translation APIs, or augmentation are used. Tokenization is parallelized across 4 CPU threads.

Hyperparameters. Table 2 summarizes our settings. Both hybrid and expert models train for 4 epochs with AdamW (Loshchilov and Hutter, 2017) and a Cosine Annealing schedule (25% linear warm-up). Full pipeline training completes on a single NVIDIA H100 (80 GB) GPU

Hyperparameter	Value
Backbones	mDeBERTa-v3-base + XLM-R-large
Quantization	4-bit NF4 (double quant.)
Precision	`bfloat16` (H100 native)
LoRA r / α (hybrid)	32 / 64
LoRA r / α (expert)	16 / 32
Epochs / Batch size	4 / 128
LR hybrid / expert	3×10^{-4} / 3×10^{-5}
Max seq. length	256 tokens
Focal γ, smooth ε	2.0, 0.05
Meta-stacker	XGBoost + Platt calibration
Threshold grid	81 pts, $t \in [0.1, 0.9]$
Hardware	NVIDIA H100 (80 GB)

Table 2: Hyperparameters for SEMANTICVECTORS.

in `bfloat16` precision. All models are loaded and fine-tuned using the HuggingFace Transformers library (Wolf et al., 2020).

Computational cost. Inference on an H100 requires ≈28 GB VRAM and ≈340 samples/sec (batch = 32, seq = 256), ≈2× the latency of a single-encoder baseline; XGBoost and Platt calibration add <1 ms/sample. The SV-BASE variant retains ≈96% of full-system F1.

Baselines. We evaluate against four configurations: (0) the **official task baseline** provided by the POLAR@SemEval-2026 organizers, a majority-class predictor (reported macro-F1 = 0.461); (1) mDeBERTa-v3-base alone (QLoRA + focal); (2) XLM-R-large alone (QLoRA + focal); (3) XLM-R-large with standard cross-entropy (reported as FULL – focal in Table 4).

Reproducibility. All experiments use random seed 42. Results reflect a single training run; variance across seeds was not evaluated due to compute constraints.

5 Results

5.1 Official Evaluation

Our system (**0.797**) substantially outperforms the official POLAR majority-class baseline (0.461) by **+0.336** points. Table 3 reports SV-FULL results: macro-F1 = **0.797**, accuracy = **0.827**. Performance peaks on morphologically consistent, script-uniform languages—Nepali (0.909), Chinese (0.886), Telugu (0.873)—within the Sinitic and Indic families. Italian (0.644) and German (0.726) prove most challenging due to implicit ironic framing.

Importantly, we observed minimal threshold overfitting; the calibrated decision boundaries generalized remarkably well to the hidden test set. This was particularly evident in critically low-resource languages, where Hausa (Dev F1 = 0.811, Test F1 = 0.803) and Odia (Dev F1 = 0.766, Test F1 = 0.771) maintained highly consistent performance across splits without degrading.

Language	Acc.	Base F1	Our F1	Rank
Amharic (amh)	**.842**	**.715**	**.780**	**7 / 30**
Arabic (arb)	.831	.796	.830	11 / 33
Bengali (ben)	.831	.853	.827	25 / 37
German (deu)	**.726**	**.671**	**.726**	**9 / 33**
English (eng)	.809	.780	.798	14 / 44
Persian (fas)	.851	.842	.803	18 / 32
Hausa (hau)	.924	.775	.803	11 / 31
Hindi (hin)	**.901**	**.738**	**.811**	**9 / 35**
Italian (ita)	**.661**	**.677**	**.644**	**10 / 32**
Khmer (khm)	**.922**	**.659**	**.755**	**3 / 31**
Myanmar (mya)	.864	.821	.860	23 / 30
Nepali (nep)	**.909**	**.880**	**.909**	**6 / 33**
Odia (ori)	.817	.777	.771	22 / 33
Punjabi (pan)	.768	.790	.768	18 / 33
Polish (pol)	.806	.724	.800	20 / 32
Russian (rus)	.805	.746	.773	22 / 31
Spanish (spa)	.782	.727	.782	11 / 36
Swahili (swa)	.780	.757	.780	16 / 31
Telugu (tel)	.873	.644	.873	20 / 33
Turkish (tur)	**.789**	**.696**	**.789**	**10 / 31**
Urdu (urd)	.807	.789	.771	21 / 35
Chinese (zho)	.886	.869	.886	21 / 33
System Avg	**.827**	**.760**	**.797**	—

Table 3: SV-FULL performance on the hidden test set. Bold rows indicate languages where the system achieved a Top-10 leaderboard rank.

Comparison with Generative LLMs. To establish a rigorous baseline against modern generative architectures, we evaluated the performance of **Llama-3-8B-Instruct** in a zero-shot setting on the official development set to ensure a fair comparison on held-out data. Using a standardized multilingual inference prompt, the LLM achieved an overall macro-F1 of 0.534. Our task-specific Siamese architecture (0.797) outperformed this generative baseline by a massive margin (+26.3 pp). This reinforces our hypothesis that while large generative models possess broad linguistic coverage, they suffer from the performance cost of instruction-following fine-tuning and lack the specialized representational depth required to reliably detect culturally-embedded rhetorical signals in low-resource settings (e.g., Llama-3 achieved only F1 = 0.281 on Hausa and F1 = 0.299 on Odia).

Low-resource language analysis. Figure 2 plots macro-F1 against training sample count on a log scale, empirically validating the degradation trend identified in our limitations. Languages below the 500-sample threshold (shown in red: hau, ori, amh, swa) exhibit a systematic *high-recall / low-precision* failure pattern—the model over-predicts the polarized class when training signal is sparse. High-resource languages show the inverse pattern (low recall, high precision; e.g., eng Rec = 0.784 vs. ita Rec = 0.467). Platt calibration partially mitigates the over-prediction bias by recalibrating pos-

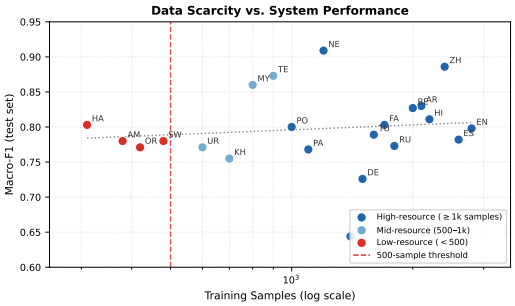
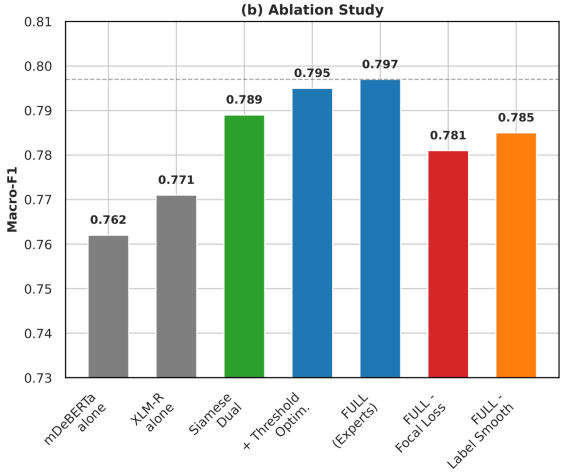


Figure 2: Data scarcity vs. system performance. Each point is one language; blue = high-resource (≥1k), light blue = mid-resource (500–1k), red = low-resource (<500). Dotted line = log-linear trend. The performance drop below the dashed 500-sample threshold is consistent and cross-family.

System Configuration	$F1_{mac}$	Δ
mDeBERTa-v3-base alone	0.762	—
XLM-R-large alone	0.771	+0.9
Siamese dual-encoder	0.789	+1.8
+ threshold optim.	0.795	+0.6
+ expert ens. (FULL)	**0.797**	+0.2
FULL − focal loss (CE)	0.781	−1.6
FULL − label smoothing	0.785	−1.2

Table 4: Ablation study on the development set. **Δ** in the upper block is relative to the previous row; **Δ** in the lower block is relative to FULL.



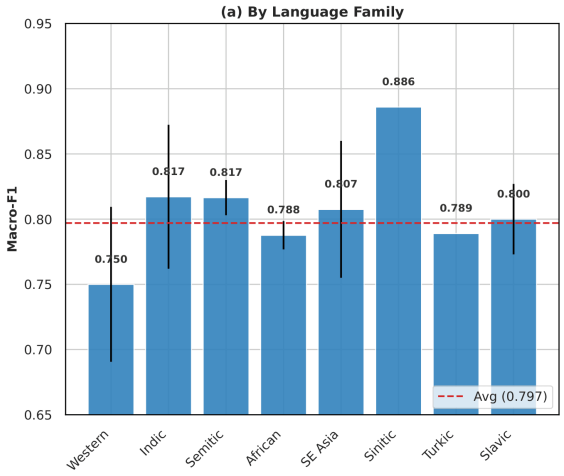
(a) Ablation study demonstrating the +1.8 pp gain.

Figure 3: System Performance Analysis.

teriors to match the dev-set class distribution, but cannot fully compensate for sparse task-specific supervision.

5.2 Ablation Study

Table 4 and Figure 3 quantify each architectural decision. The Siamese dual-encoder delivers the largest gain (+1.8 pp), confirming that complementary inductive biases—mDeBERTa's syntactic precision and XLM-R's broad semantic coverage—outperform scaling a single encoder. Focal loss



(a) Macro-F1 by language family.

Figure 4: Development Trajectory. Strong Sinitic/Indic transfer across language families.

is also critical: ablating it costs −1.6 pp, as its hard-example mining concentrates model capacity on subtly framed rhetorical instances that standard cross-entropy treats as equally weighted.

Development Trajectory & Language Families. Figure 4 shows performance stratified by language family, with Sinitic (0.886) and Indic (avg. 0.817) highest and Western (avg. 0.750) showing the greatest variance.

6 Error Analysis and Robustness

Error Typologies. We systematically expanded our manual inspection to 300 randomly sampled errors from the development set to ensure robust typological coverage. The distribution reveals that false negatives (54.3%) slightly outpace false positives (45.7%), confirming a residual conservative bias.

Implicit Rhetorical Framing (False Negatives): The system struggles significantly when polarization relies on cultural dog-whistling or structural irony rather than explicit hostility. This was highly visible in Italian and German, where texts utilizing rhetorical questions to mock out-groups were classified as neutral due to their lack of toxic vocabulary. Similarly, low-resource languages like Hausa suffered from missing sociopolitical context.

Topical Bias (False Positives): Conversely, false positives frequently occurred when neutral, objective journalism covered highly polarized political topics. The encoders occasionally over-indexed on specific divisive terminology (e.g., "Anti-Woke" in German analytical texts), failing to recognize the neutral syntactic envelope surrounding those tokens. Table 5 provides representative real-world examples drawn from our 300-sample analysis.

Lang.	Text Extract (Translated/Paraphrased)	Gold	Pred
eng	"we build a wall around the red states…" *(Implicit framing / sarcasm)*	1	0
deu	"Reaktionäres Anti-Grünen / Anti-Woke…" *(Analytical discussion; topical bias)*	0	1
hau	"…the fulanis are the real almajiri…" *(Low-resource ethnic dog-whistling)*	1	0

Table 5: Representative error cases from the 300-sample analysis using SV-FULL.

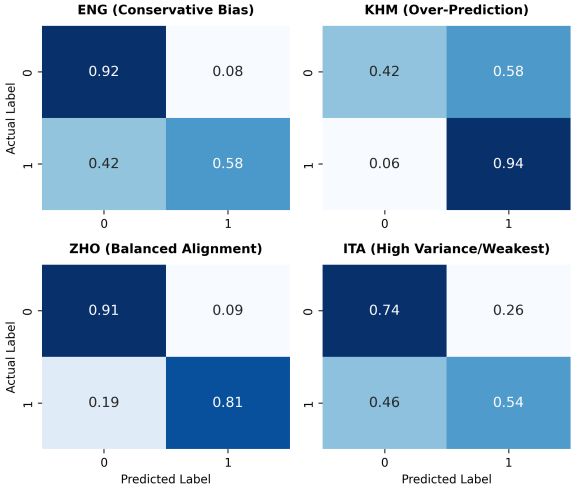


Figure 5: Normalized confusion matrices across all 22 languages (dev set). 0 = neutral, 1 = polarized. Three patterns: *conservative bias* (eng, deu), *over-prediction* (khm, amh), *balanced* (zho, mya).

Per-Language Confusion Analysis. Figure 5 presents normalized confusion matrices for all 22 languages, revealing three systematic error patterns. *Conservative bias* (high TNR, high FNR) dominates high-resource Western languages: English (TNR = 0.92, FNR = 0.42) defaults to "neutral" when explicit divisive vocabulary is absent. *Over-prediction bias* (high FPR) appears in low-resource languages: Khmer misclassifies 58% of neutral texts as polarized. A *balanced* regime appears in Sinitic and Indic families (Chinese: TNR = 0.91, TPR = 0.81). These patterns suggest that the conservative vs. over-prediction axis correlates with training data volume rather than language family per se.

Robustness to Sequence Length. The Siamese architecture is highly robust to brief sequences, achieving macro-F1 ∼0.82 on ultra-short texts (0–15 words, n=2104), comparable to extended texts (51+ words, n=87, $\Delta \leq 0.04$; note small sample size). This contradicts the hypothesis that short texts create a representational bottleneck, and is attributable to the dense $\mathbb{R}^{1792}$ joint representation: the dual-encoder embeds a separable polarization signal from minimal surface text without requiring extended contextual padding.

7 Conclusion

SEMANTICVECTORS achieves macro-F1 = 0.797 and accuracy = 0.827 across 22 languages on POLAR@SemEval-2026 Task 9. The Siamese dual-encoder (+1.8 pp) and focal loss as a hard-example miner (−1.6 pp when ablated) are the core contributions. The XGBoost meta-stacker—leveraging prediction confidence, Shannon entropy, and token count as reliability signals—provides better-calibrated fusion than soft voting, and fusing two complementary tokenizers (SentencePiece + BPE) reduces effective OOV rates for morphologically rich low-resource scripts. Negative experiments with RWKV, Mamba, and BitNet suggest that representational breadth dominates efficiency in the 22-language setting. Error analysis identifies two tractable failure modes: conservative bias in high-resource Western languages (eng FNR = 0.42), and over-prediction bias in low-resource settings (khm FPR = 0.58) that correlates with training data volume as shown empirically in Figure 2. Future work targets cross-lingual distillation, rhetorical-question fine-tuning, and language-family-conditioned augmentation for low-resource settings.

Limitations

Performance degrades below 500 training examples, and low-resource languages show systematic over-prediction bias that Platt calibration partially mitigates but cannot resolve without additional training data. Operating two large encoders doubles inference latency (≈2×) and VRAM requirements vs. a single-model baseline.

Ethics Statement

Our system detects polarization; it does not generate or amplify it. All training data are from the official task organizers. Automated classifiers carry misuse risks (censorship, profiling) and require human oversight. The performance gap on low-resource languages may disproportionately affect speakers of those languages.

Acknowledgments

We thank the POLAR@SemEval-2026 Task 9 organizers for datasets and evaluation infrastructure. We acknowledge Google Colab, Kaggle, and Lightning AI for GPU resources during early development. Finally, we acknowledge the use of Gemini, Claude and ChatGPT strictly for copy-editing and LaTeX formatting assistance during the preparation of this manuscript.

References

David Adelani. xlm-roberta-base-finetuned-swahili, 2022. URL https://huggingface.co/Davlan/xlm-roberta-base-finetuned-swahili. Hugging Face Model Hub.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.

Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *Proceedings of the 28th international conference on computational linguistics*, pages 6788–6796, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451, 2020.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2): 3, 2022.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille, 2015.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

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. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, 2026a.

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, 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. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization, 2026b. URL https://arxiv.org/abs/2505.20624.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 14048–14077, 2023.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Stefan Schweter. Italian bert and electra models. *Zenodo*, 2020.

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. Bitnet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.