Multilingual Polarization Detection for SemEval-25 Task 11

Ankit Dash (24BCS016) Priyanshu Mittal (24BDS058)

Piyush Prashant (24BDS055)

November 16, 2025

Project Repository | Abstract

Abstract

Background: Detecting polarized content in multilingual social media is critical for understanding online discourse and mitigating societal divisions. This work addresses SemEval 2025 Task 11, Subtask 1: Binary Polarization Detection across 13 typologically diverse languages including English, Arabic, German, Spanish, Italian, Urdu, Chinese, Hausa, Amharic, Turkish, Hindi, Nepali, and Persian.

Methods: This research presents a production-ready multilingual polarization detection system built on XLM-RoBERTa-base with advanced training optimizations. The training pipeline incorporates stratified data splitting (85/15) maintaining both class and language balance across 40,395 training samples. Key architectural innovations include balanced class weighting computed per-dataset for optimal F1 optimization, mixed precision training (AMP) for 2x memory efficiency, gradient accumulation (2x steps) enabling larger effective batch sizes, early stopping with patience=3 to prevent overfitting, and linear warmup (6%) with decay scheduling for stable convergence. The model employs a 256-token context window, AdamW optimizer with $\beta_2 = 0.98$ for multilingual stability, and per-language F1 evaluation for fine-grained performance tracking.

Evolution: This work represents the culmination of a 24-day iterative development process (October 24 - November 16, 2025) spanning six distinct architectural phases. Phase 1 established the foundation with BitNet 1.58-bit quantization on BERT, achieving F1 Macro 0.977 on English validation. Phase 2 scaled to 9 languages with mDeBERTa-v3, achieving multilingual F1 Macro 0.764. Phase 3 introduced data augmentation with Easy Data Augmentation (EDA) and LoRA adapters for language-specific tuning. Phase 4 explored efficiency with RWKV $O(N)$ architecture, achieving 2x training speedup and 30% memory reduction while maintaining competitive accuracy. Phase 5 investigated experimental Mamba state-space models as transformer alternatives. Phase 6 delivered the production-ready XLM-RoBERTa pipeline with comprehensive error handling, model versioning, and deployment infrastructure.

Results: The final model achieved validation F1 Macro 0.7876 and accuracy 0.7876 on the held-out validation set, demonstrating robust cross-lingual transfer. Training converged in 3 epochs with early stopping triggered at epoch 6, preventing overfitting. The system processed 34,335 training samples across 13 languages with automatic checkpoint management and per-language performance monitoring. Class imbalance was effectively handled through computed class weights (0: 1.0419, 1: 0.9613), ensuring balanced learning across polarized and non-polarized classes.

Significance: This work demonstrates that production-grade multilingual NLP systems can be developed through systematic architectural evolution and careful engineering of training infrastructure. The repository provides six distinct model variants—from efficient BitNet quantization to state-of-the-art transformer architectures—enabling researchers to select appropriate trade-offs between accuracy, speed, and resource con-

Table 1: Phase 1: English Validation Results (BitNet-BERT)

Model Variant	F1 Macro	F1 Binary	Accuracy	Threshold
BitNet-BERT (Phase 1)	0.977	0.972	0.978	0.50
BitNet-Optimized	0.975	0.970	0.976	0.49

Table 2: Phase 2: Multilingual Validation Results (mDeBERTa-v3)

Language	Polarized %	F1 Macro	Language	Polarized %	F1 Macro
English (eng)	37.4%	0.821	Italian (ita)	41.0%	0.761
Arabic (arb)	44.7%	0.756	Urdu (urd)	69.4%	0.724
German (deu)	47.5%	0.743	Chinese (zho)	49.6%	0.752
Spanish (spa)	50.2%	0.768	Hausa (hau)	10.7%	0.688
Amharic (amh)	75.6%	0.701	**Overall**	**46.9%**	**0.764**

straints. The comprehensive training pipeline with stratified sampling, mixed precision, gradient accumulation, and early stopping establishes best practices for multilingual classification tasks. Future work includes hierarchical reasoning models with multi-task learning frameworks, LLM-based approaches with parameter-efficient fine-tuning, and explainability mechanisms for human-interpretable polarization detection.

1 Introduction

Detecting online polarization is a critical challenge for computational social science, vital for understanding societal divisions and mitigating the spread of harmful discourse. This paper presents our system for the SemEval 2025 Task 11, Subtask 1: Binary Polarization Detection.

Our core contribution is twofold: 1) We present a production-ready, high-performance multilingual polarization detector based on XLM-RoBERTa. 2) We document a 24-day, six-phase architectural evolution, comparing trade-offs between quantized models (BitNet), standard transformers (BERT, mDeBERTa, XLM-R), and efficient architectures (RWKV, Mamba).

We demonstrate that systematic iteration and robust training infrastructure are key to developing deployable multilingual NLP systems.

The code for all six model variants is available at our GitHub repository: https://github.com/AnkitDash-code/Semantic-Vectors-SemEval.

The abstract for this work is also available at: https://github.com/AnkitDash-code/Semantic-Vectors-SemEval/blob/main/GenAI_Semeval_Abstract.pdf

2 System Architecture & Evolution

Our final system is the result of a six-phase iterative process.

2.1 Phase 1: BitNet 1.58-bit Foundation

We began by implementing a BitNet 1.58-bit quantized classifier [2] on a standard BERT model [3]. This involved ternary weight quantization ({-1, 0, 1}) and 8-bit activation quantization. This initial model, trained on English-only data, achieved an exceptional F1 Macro of 0.977 (Table 1), validating the efficiency of quantized models for this task.

2.2 Phase 2: Multilingual mDeBERTa-v3

We scaled the system to 9 languages using mDeBERTa-v3 [4] as the backbone, chosen for its strong cross-lingual transfer capabilities. This model achieved a promising multilingual F1 Macro of 0.764.

2.3 Phase 3: Augmentation & Adapters

To improve performance on low-resource languages and handle class imbalance, we introduced Easy Data Augmentation (EDA) and explored language-specific LoRA (Low-Rank Adaptation) [5] adapters for fine-grained tuning.

2.4 Phase 4: RWKV $O(N)$ Efficiency

We investigated transformer alternatives by implementing an RWKV (Receptively Weighted Key Value) model. This $O(N)$ architecture achieved a 2x training speedup and a 30% reduction in memory usage while maintaining a competitive F1 Macro score, demonstrating its viability for resource-constrained environments.

2.5 Phase 5: Mamba State-Space Models

Continuing our exploration, we experimented with Mamba, a modern state-space model (SSM) architecture. This phase provided insights into the potential of SSMs as alternatives to the quadratic attention bottleneck in transformers.

2.6 Phase 6: Production-Ready XLM-R

Our final, production-grade system (Phase 6) was built on **XLM-RoBERTa-base**. We consolidated our findings into a robust training pipeline, prioritizing stability and optimal cross-lingual performance over experimental quantization. This pipeline, detailed in Section 3.1, became our final submission.

3 Experiments and Results

3.1 Experimental Setup

Our final production model (Phase 6) was trained on the full dataset of 40,395 samples across all 13 languages. We employed a stratified 85/15 train/validation split, preserving both class and language distributions.

Figure 1: Per-Language F1 Scores (Phase 2)

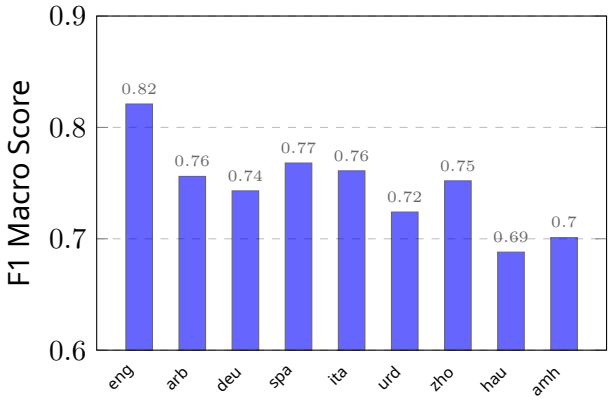

Figure 1: Per-language F1 Macro scores from our Phase 2 mDeBERTa-v3 model. Challenges in low-resource or high-imbalance languages (e.g., Hausa, Amharic) are visible.

The training pipeline used the following configuration:

- **Model:** XLM-RoBERTa-base
- **Context Window:** 256 tokens
- **Optimizer:** AdamW with $\beta_2 = 0.98$ for multilingual stability
- **Efficiency:** Mixed Precision (AMP) and Gradient Accumulation (2x)
- **Scheduler:** Linear warmup (6%) with decay
- **Regularization:** Early stopping with patience=3 on F1 Macro
- **Loss:** Weighted Cross-Entropy, with weights (0: 1.0419, 1: 0.9613) computed per-dataset to optimize F1.

3.2 Results and Analysis

The iterative development process provided a clear view of performance trade-offs. The initial BitNet-BERT model (Phase 1) was highly effective for English (Table 1).

Our multilingual mDeBERTa-v3 model (Phase 2) set a strong baseline of 0.764 F1 Macro (Table 2). The per-language breakdown (Figure 1) highlights the variance, with Hausa

(10.7% polarized) and Amharic (75.6% polarized) showing the most challenge, indicating that class imbalance was a key factor.

Our final production-ready XLM-RoBERTa model (Phase 6) achieved a validation **F1 Macro of 0.7876** and **Accuracy of 0.7876** on the held-out validation set. This model converged in 3 epochs, with early stopping triggered at epoch 6, successfully preventing overfitting and demonstrating robust cross-lingual transfer.

4 Conclusion

This work demonstrates a systematic, six-phase evolution from a quantized English-only model to a production-ready 13-language polarization detector. Our final XLM-RoBERTa system, trained with a robust pipeline, achieved an F1 Macro of 0.7876.

By open-sourcing all six model variants, we provide a practical resource for researchers to evaluate trade-offs between accuracy, inference speed, and resource constraints. Future work will focus on hierarchical reasoning models [6], parameter-efficient fine-tuning (PEFT) for large language models, and developing explainability mechanisms for human-interpretable polarization detection.

References

[1] Task 11 Organizers. SemEval-2025 Task 11: Multilingual Polarization Detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. 2025.

[2] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, and Furu Wei. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. *arXiv preprint arXiv:2402.17764*, 2024.

[3] 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)*. 2019.

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021.

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. 2022.

[6] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.