PolarMind at SemEval-2026 Task 9: Leveraging LaBSE with Progressive Curriculum Learning for Multicultural Polarization

Sandeep Kumar[†] **Mothish**[♣] **Sachin Sundar**[♣]

[†]Indian Institute of Technology, Kharagpur,
[♣]Indian Institute of Technology, Madras
sandeepkumar.24@kgpian.iitkgp.ac.in
{cs24b033, mm24b013}@smail.iitm.ac.in

Abstract

Detecting online polarization remains a critical challenge, particularly in multilingual and multicultural contexts where intergroup hostility is prevalent. The problem is particularly challenging due to the data scarcity for these tasks in the low-resource languages. Identifying such phenomena has become an active area of research and is addressed in SemEval-2026 Task 9: Multilingual, Multicultural Online Polarization Detection. To address this problem we propose an architecture that leverages LaBSE embeddings—an unconventional choice typically reserved for retrieval tasks—to obtain strong cross-lingual learning which enhances scores in low-resource language by a score up to 0.2 macro F1. Furthermore, we provide a comprehensive ablation study evaluating the performance of diverse encoder models in the Qwen model family within a retrieval-based prompting framework.[1]

1 Introduction and Related Work

In the contemporary digital landscape, social networks have emerged as a primary medium of information exchange and a source through which people belonging to diverse populations interact. However, it has become common to see inter-group hostility and antagonistic viewpoints expressed across various platforms.

Digital platforms often amplify extreme views, degrading the quality of online debate (Iandoli et al., 2021). Because of this, identifying polarized speech is now a critical challenge for the research community. Polarized text can be broadly classified into 4 main categories: political, religious, racial/ethnic, and gender/sexual. The major challenge for building reliable systems that detect these is that the text can be polarized due to a single harsh word or even due to the overall opinion expressed in the comment. Polarization is subject to different ethnic groups and cultures (Kannen et al., 2025), i.e., a completely unpolarised sentence in English when translated to Urdu might be polarized(Zafar et al., 2025). Consequently, the optimal method to approach this problem would be to design specialized systems for each language. However, most of the state-of-the-art models today exhibit large performance gaps between English and low-resource languages(Verma et al., 2025; Saji et al., 2025). Also, with a limited amount of training data, it becomes necessary for our systems to group languages together in-order to gain better overall performance through cross-lingual learning. To address these challenges, this paper presents our approach for SemEval 2026 task 9:"Multilingual Text Classification Challenge"(Naseem et al., 2026a). We participate in 2 subtasks each with 22 languages in each subtask (Naseem et al., 2026b).

- **Subtask 1:** Polarization detection (binary classification)
- **Subtask 2:** Polarization type detection with 5 classes (Political, Racial, Gender, Ethnic, Other)

In this work, we propose a unique architecture using layer fusion and hybrid pooling (attention pooling + mean pooling) in LaBSE embeddings along with Proxy-Guided Curriculum Learning to get enhanced results. In a detailed ablation study, we compare RemBERT with and without continual pretraining, EuroBERT — a state-of-the-art model with more recent knowledge — and Qwen-2.5-14B using in-context learning with MMR-reranked sentence retrieval. For more details regarding our model architecture and the proxy-guided curriculum learning strategy, refer to Section 2.

Contributions

- LaBSE-based architecture with weighted layer aggregation and hybrid pooling for cross-lingual polarization detection.
- Proxy-guided curriculum learning to address

[1]Our code will be soon available at https://github.com/carrycurious/PolarMind

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3232–3239

multilingual data imbalance.

- Comparative evaluation of encoder models and LLMs (Qwen-2.5).
- Empirical analysis of LLM and phoneme prompting limitations in low-resource settings.

2 System Architecture and Ablations

This section details our primary system submitted for the final shared task evaluation: a customized LaBSE architecture. To justify our design choices, we also present an ablation study comparing this core system against alternative models.

2.1 Core Architecture: LaBSE with Hybrid Pooling

We employ LaBSE (Feng et al., 2022), a bi-encoder architecture designed to map multilingual sentences with equivalent semantics into a shared embedding space. Unlike standard multilingual encoders like mBERT or XLM-R, LaBSE is fundamentally optimized to map semantically similar cross-lingual sentences to the exact same vector space. We hypothesize that this strict alignment is uniquely beneficial for multicultural polarization, allowing the model to project scarce low-resource polarized phrases into the robust semantic neighborhoods established by high-resource languages. The main components of our architecture are as follows:

Weighted Layer Aggregation: This model aggregates the final four layers of the models with learnable parameters (w_1, w_2, w_3, w_4), where (w_1, w_2, w_3) are initialized to be much smaller $(\sim 0.2w_1)$.

$$H_{final} = \sum_{i=1}^{4} w_i \cdot L_{n-4+i} \tag{1}$$

Hybrid pooling: We propose a hybrid pooling mechanism that computes a linear combination of global mean pooling and additive-attention (Bahdanau et al., 2016) pooling. This approach is designed to prioritize and highlight the most significant tokens.

The global semantic context is captured by calculating the arithmetic mean of the hidden states:

$$\mathbf{e}_{\text{mean}} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{h}_i \tag{2}$$

where T denotes the sequence length and $\mathbf{h}_i$ represents the i-th hidden vector.

To emphasize high-impact tokens, we employ an additive attention mechanism:

$$u_i = \tanh(\mathbf{W}\mathbf{h}_i + \mathbf{b}) \tag{3}$$

$$\alpha_i = \frac{\exp(\mathbf{w}^\top u_i)}{\sum_{j=1}^{T} \exp(\mathbf{w}^\top u_j)} \tag{4}$$

$$\mathbf{e}_{\text{attn}} = \sum_{i=1}^{T} \alpha_i \mathbf{h}_i \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{k \times d}$ and $\mathbf{b}$ are learnable parameters, and $\mathbf{w}^\top$ is the context vector.

The final representation is a weighted sum of the global and salient features, modulated by the hyperparameter λ:

$$\mathbf{e}_{\text{hybrid}} = \mathbf{e}_{\text{mean}} + \lambda \mathbf{e}_{\text{attn}} \tag{6}$$

The mean pooling component (e_{mean}) stabilizes the representation by capturing overall semantic context, while the additive attention mechanism (e_{attn}) emphasizes high-impact tokens relevant to classification. In Subtask 2, each classification head maintains an independent attention mechanism, ensuring that token importance for one polarization type (e.g., gender) does not conflate with another (e.g., ethnic). The final representation $\mathbf{e}_{\text{hybrid}}$ is used for classification.

2.2 Ablation Study: Alternative Models

To evaluate the effectiveness of our core LaBSE architecture, we explored alternative models spanning both encoder-only and large language model (LLM) paradigms.

ii) RemBERT with continual pretraining: To further extend our study, we leverage RemBERT (Chung et al., 2020), a high-capacity multilingual encoder, for the classification task. We perform continual pre-training on our dataset to adapt the model and make it aware of the recent events, as the model was released in 2020. For continual pretraining, we create a corpus using all the data in Subtask 3 (which we did not take part in) and use this for pretraining.

iii) EuroBERT-210m: EuroBERT (Boizard et al., 2025) is a state-of-the-art encoder model, optimized for English and european languages. EuroBERT (2024) has a more recent knowledge cut-off and hence has seen more information based on recent global events. We use this model to analyze the effect of recent knowledge.

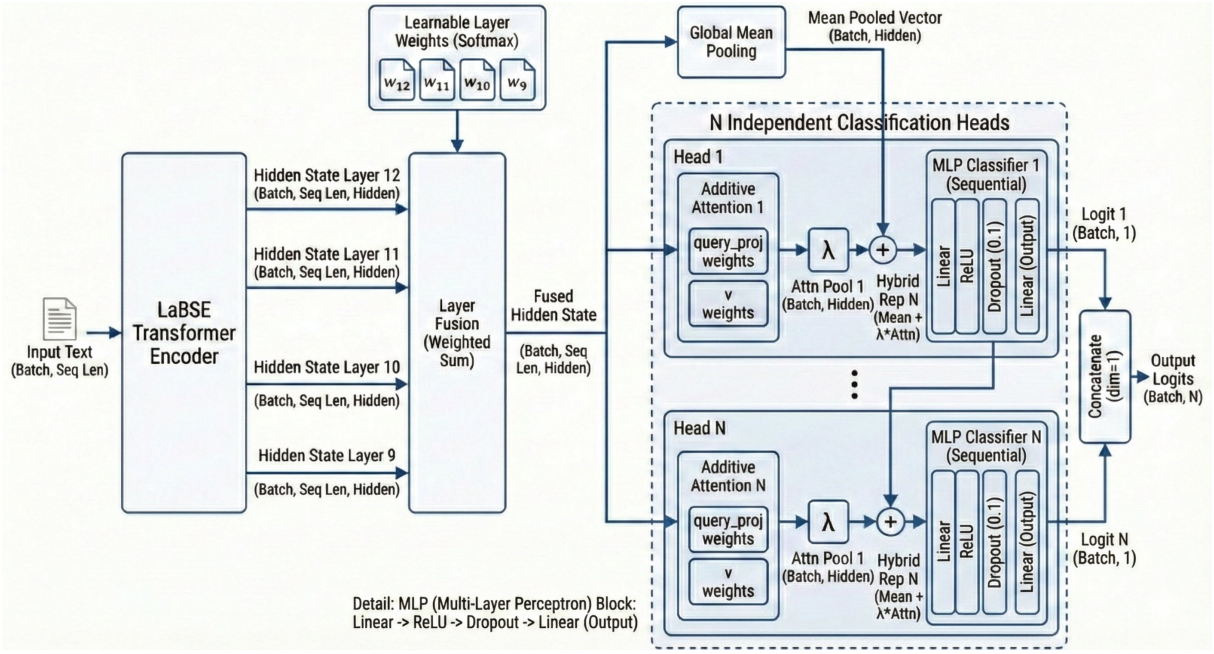

Figure 1: Proposed Hybrid Contextual Pooling architecture.

iv) QWEN-2.5 with retrieval using phonemes: Recent advancements have shown that modern LLMs with optimized prompting, exhibit comparable performance with traditional encoder models in classification tasks(Wang et al., 2024). We use few-shot prompting (Brown et al., 2020) with the following strategies:

Multilingual Sentence Retrieval We use LaBSE as a retriever to find the most(top k) semantically similar sentences from our training data for a given input sentence. This allows the model to see and understand how similar polarization manifests across different languages and cultural contexts before making a final prediction.

Maximal Marginal Relevance (MMR) for Diversity To ensure that retrieved examples are both relevant and diverse, we implement the Maximal Marginal Relevance (MMR) algorithm (Carbonell and Goldstein, 1998). This algorithm ensures that prompt redundancy,i.e, makes sure that prompts are similar to the given query Q but yet diverse. For a given query Q, MMR iteratively selects a sentence D_i from candidate set R as follows:

$$\text{MMR} = \arg \max_{D_i \in R \setminus S} \big[\alpha \cdot \text{Sim}_1(D_i, Q) - (1-\alpha) \cdot \max_{D_j \in S} \text{Sim}_2(D_i, D_j)\big] \quad (7)$$

where $\text{Sim}_1(D_i, Q)$ is the cosine similarity between candidate D_i and query Q (relevance), $\max \text{Sim}_2(D_i, D_j)$ is the maximum similarity to already selected sentences S (diversity), and α=0.7 balances the two.

Few-Shot Prompting with Qwen-2.5 The top k diverse sentences selected via MMR are formatted into a structured prompt as shown in the Appendix B. Using this retrieved context, Qwen-2.5-14B (Qwen et al., 2025) performs in-context learning to categorize the target input, leveraging the diverse examples to better capture the nuances of the classification task.

Phonemes based prompting Nguyen et al. (2025) has shown that introducing phoneme information to LLMs along with raw scripts (*text+ipa*)

Model	Strategy	Pooling	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
LaBSE	mixed	ours	0.723[.77]	0.783[.81]	**0.818**[.82]	**0.735**[.69]	0.804[.80]	**0.883**[.81]	**0.838**[.82]	**0.843**[.77]	0.647[.64]	0.697[.71]	**0.923**[.86]
	separate	ours	**0.753**	**0.812**	0.801	0.627	0.798	0.765	0.798	0.835	0.618	**0.719**	0.857
Vanilla RemBERT	mixed	Cls	0.698	0.792	0.796	0.709	0.741	0.821	0.693	0.842	**0.663**	0.652	0.852
RemBERT (cont.)	mixed	Cls	0.679	0.784	0.811	0.674	0.768	0.834	0.710	0.838	0.661	0.638	0.876
EuroBERT	mixed	Cls	—	—	—	0.729	**0.818**	—	—	—	0.586	—	—
Qwen-2.5-Instruct-14b	0 shots	text	0.598	0.765	0.732	0.728	0.741	0.649	0.582	0.643	0.485	0.127	0.548
	7 shots	text	0.685	0.740	0.698	0.687	0.764	0.698	0.601	0.689	0.517	0.527	0.765
	7 shots	text+ipa	0.624	0.756	0.687	0.719	—	0.746	0.724	0.658	0.565	—	0.793

Model	Strategy	Pooling	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
LaBSE	mixed	ours	**0.886**$_{[.90]}$	0.771$_{[.79]}$	**0.819**$_{[.76]}$	0.768$_{[.77]}$	**0.778**$_{[.75]}$	0.685$_{[.75]}$	**0.798**$_{[.79]}$	0.881$_{[.87]}$	0.799$_{[.76]}$	**0.727**$_{[.76]}$	**0.878**$_{[.85]}$
	separate	ours	0.859	**0.791**	0.789	0.758	0.768	0.659	0.754	0.885	0.776	0.683	0.874
Vanilla RemBERT	mixed	Cls	0.850	0.701	0.810	**0.804**	0.775	**0.696**	0.778	0.908	0.790	0.693	0.867
RemBERT (cont.)	mixed	Cls	0.837	0.728	0.803	0.798	0.757	0.672	0.758	**0.912**	**0.804**	0.723	0.852
EuroBERT	mixed	Cls	—	—	—	0.750	—	0.692	—	—	—	—	—
Qwen-2.5-14b	0 shots	text	0.736	0.685	0.578	0.717	0.636	0.643	0.569	0.576	0.751	0.659	0.794
	7 shots	text	0.725	0.692	0.767	0.735	0.578	0.623	0.690	0.713	0.758	0.689	0.864
	7 shots	text+ipa	0.745	—	—	0.718	0.658	0.672	0.728	0.735	0.748	0.719	0.856

Table 1: Macro F1 scores for Subtask 1 (Dev and Official Test sets). LaBSE subscripts denote official Test results. For Qwen, "Strategy" = shot count, "Pooling" = prompting technique. **Bold** = best per language.

improves performance in various downstream tasks. Following this approach, we utilize the *Epitran* (Mortensen et al., 2018) library to convert text from each language into phonemes for further analysis.

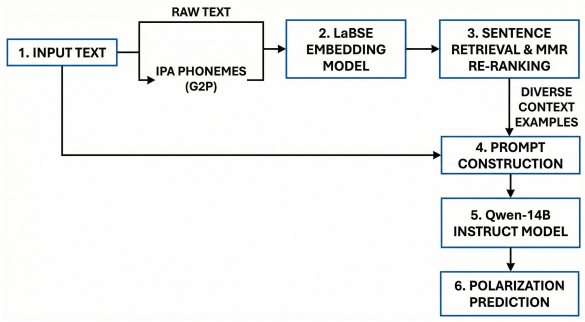

Figure 2: Pipeline that integrates phonemes

3 Experimental setup

3.1 Dataset

For both evaluation and training of our data, we use the dataset provided by organisers of Semeval task 9 Organizers (Naseem et al., 2026b). For all subsequent analyses, we report performance metrics as detailed in Table 1 on development set only.

Proxy guided Curriculum Learning. We employ a curriculum learning strategy for training all encoder models described in Section 2.To implement curriculum learning, we first categorize the dataset by difficulty using XLM-RoBERTa-base (Conneau et al., 2020), a smaller model for our purpose. We then rank the entire data based on the following equation:

$$R(x) = \alpha \cdot \mathrm{P}(y|x) + \beta \cdot \frac{1}{\mathrm{len}(x)}$$

where $\mathrm{P}(y|x)$ is the proxy confidence and $\beta/\mathrm{len}(x)$ promotes lower-fertility high-resource languages to higher ranks, partitioning data into *easy*, *medium*, and *hard* training groups.

3.2 Training setup

Encoder models Encoder models are trained with the same strategy for both subtasks, differing only in the number of classification heads. Each model undergoes four epochs following the curriculum: easy samples first, then easy+medium, then medium+hard, and finally the full dataset. We use AdamW ($lr = 2e{-}5$, weight decay 0.01) with a Cosine Annealing Scheduler (Loshchilov and Hutter, 2016), trained on a single NVIDIA L4 GPU using HuggingFace `transformers` (HuggingFace et al., 2020). Full hyperparameters are in Appendix A.

Language-Specific Decision Calibration For the multi-label task, we apply sigmoid activation and calibrate language- and class-specific thresholds $t_{\ell,c}$ on a 20% training split to maximize Macro-F1:

$$t^*_{\ell,c} = \arg\max_t \mathrm{MacroF1}_{\ell,c}(t) \tag{8}$$

4 Results

Overall, we evaluate our proposed methods and their variants on the development set to identify the optimal model configuration for each language (indicated in bold in Table 1 and Table 2). Our performance in low-resource languages is highly competitive on the official leaderboard; we achieved top-10 rankings in five languages in Subtask 1. In Subtask 2, we secured top-5 placements for five languages and top-10 rankings for eleven languages globally.

4.1 Subtask 1

Development set Our proposed LaBSE-based architecture with hybrid context pooling consistently outperformed the official `POLAR baseline` (fine-tuned LaBSE-base) by an average of **0.04 Macro F1 points** across all 22 languages. On the official SemEval-2026 leaderboard, our system demonstrated strong performance across low-resource languages, particularly in low-resource scripts. For

Model	Strategy	Pooling	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
LaBSE	mixed	ours	**0.469**$_{[.63]}$	**0.564**$_{[.59]}$	0.350$_{[.34]}$	**0.526**$_{[.55]}$	0.430$_{[.48]}$	0.551$_{[.58]}$	0.234$_{[.41]}$	**0.755**$_{[.77]}$	0.371$_{[.31]}$	**0.645**$_{[.61]}$	**0.545**$_{[.70]}$
Qwen-2.5-14b	7 shots	text	0.423	0.498	0.442	0.501	0.497	**0.569**	**0.477**	0.547	0.537	0.574	0.457
	7 shots	text+ipa	0.417	0.512	**0.521**	0.513	**0.499**	0.562	0.472	0.578	**0.541**	0.602	0.468

Model	Strategy	Pooling	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
LaBSE	mixed	ours	**0.758**$_{[.77]}$	0.501$_{[.51]}$	0.397$_{[.42]}$	0.534$_{[.51]}$	0.469$_{[.49]}$	**0.607**$_{[.63]}$	0.455$_{[.47]}$	0.463$_{[.44]}$	**0.595**$_{[.56]}$	**0.754**$_{[.76]}$	0.717$_{[.73]}$
Qwen-2.5-14b	7 shots	text	0.578	0.593	0.565	**0.563**	**0.553**	0.552	**0.548**	**0.540**	0.539	0.535	0.511
	7 shots	text+ipa	0.592	**0.643**	**0.609**	0.524	0.513	0.568	0.531	0.517	0.483	0.562	**0.793**

Table 2: Macro F1 scores for Subtask 2 on Development and Official Test set. Bolding indicates the highest value per column. Same guidelines as Table 1

instance, in Amharic (amh), we achieved a score of **0.771**, a **+5.6% improvement** over the baseline.

While EUROBERT (210M) performed at par in specific European contexts—surpassing our model in English and Italian by approximately 0.01 Macro F1—this is likely due to its more recent knowledge cut-off tailored. However, the marginal gains (**+0.005 Macro F1**) observed from continual pre-training (REMBERT-CONT) suggest that our hybrid pooling strategy provides a more significant performance boost than model scaling alone. For the decoder model (Qwen-2.5), performance is not competitive with encoder models, likely due to tokenization mismatches across scripts and weaker cross-lingual alignment compared to LaBSE. Phonetic information does not significantly improve performance. We hypothesize that the phonetic layer offers redundant information for this specific task, as the model's internal representations of raw text already encompass the necessary cultural nuances, while the IPA transformation potentially filters out emoji usage and script-specific markers that carry significant polarized weight.

4.2 Subtask 2

Development set As shown in Table 2, we focus on our LaBSE-based model as it consistently outperforms other baselines (RemBERT and EuroBERT). Language-specific threshold calibration improved the average Macro-F1 on the development set from 0.5216 to 0.5441, an absolute gain of +2.25%. These gains are most evident in languages with skewed class distributions, emphasizing the sensitivity of multi-label polarization detection to threshold selection.

Interestingly, Qwen (7-shot) performs on par with our encoder model. While the encoder struggles to generalize sparse patterns (e.g., gender or ethnicity) from limited samples, Qwen leverages its prior knowledge via few-shot prompting. Finally, while our phoneme-based approach yields minor improvements, the current results do not provide a clear pattern or conclusive evidence that modern LLMs benefit from phonetic prompting for this task.

5 Conclusion

We evaluate our proposed methods and their variants on the development set to identify the optimal model configuration for each language (indicated in bold in Table 1 and Table 2). Our performance in low-resource languages is highly competitive on the official leaderboard; we achieved top-10 rankings for five languages in Subtask 1. In Subtask 2, we secured top-5 placements for five languages and top-10 rankings for eleven languages globally.

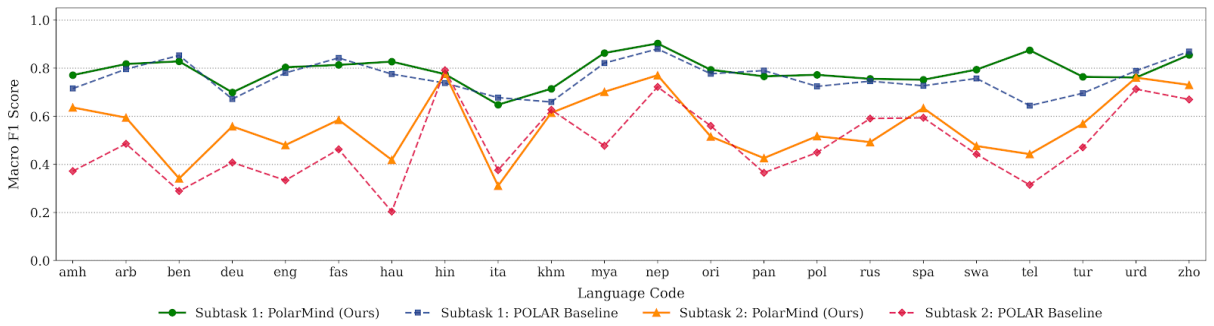

Figure 3: Comparison with POLAR baseline

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *Preprint*, arXiv:2503.05500.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared S Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *Preprint*, arXiv:2010.12821.

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, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

HuggingFace, T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Transformers - state-of-the-art machine learning for pytorch, tensorflow, and jax.

Luca Iandoli, Simonetta Primario, and Giuseppe Zollo. 2021. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*, 170:120924.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2025. Beyond aesthetics: Cultural competence in text-to-image models. *Preprint*, arXiv:2407.06863.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Hoang H Nguyen, Khyati Mahajan, Vikas Yadav, Julian Salazar, Philip S. Yu, Masoud Hashemi, and Rishabh Maheshwary. 2025. Prompting with phonemes: Enhancing llms' multilinguality for non-latin script languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 11975–11994. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2025. Romanlens: The role of latent romanization in multilinguality in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*,

page 26410–26429. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. *Preprint*, arXiv:2411.02538.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is chatgpt a good sentiment analyzer? a preliminary study. *Preprint*, arXiv:2304.04339.

Shanza Zafar, Prof. Dr. Ijaz Asghar, Dr. Naveed Nawaz, and Prof. Dr. Hafiz Ahmad Bilal. 2025. Investigating political ideology in english and urdu newspaper editorials a corpus-based comparative study. *Journal of Applied Linguistics and TESOL (JALT)*, 8(4).

A Hyperparameters

All models were trained on an NVIDIA L4 GPU. The encoder backbones were **LaBSE** and **RemBERT**. The **Qwen** model was used only for evaluation (no fine-tuning).

Training Setup

Hyperparameter	Value
Optimizer	AdamW ($\epsilon = 1 \times 10^{-5}$)
Batch Size	32
Scheduler	Cosine decay
Total Steps	len(dataloader) $\times$ 4

Subtask 1 used Cross-Entropy loss, while Subtask 2 used weighted inverse-frequency Cross-Entropy.

Fusion pooling used $\lambda = 0.2$ (Subtask 1) and $\lambda = 0.5$ (Subtask 2). For LaBSE layer weighting: $W_1 = W_2 = W_3 = 0.2$, $W_4 = 1$.

Parameter	LaBSE	RemBERT
Architecture Config		
Max Token Length	512	512
Hidden Size	768	1152
Encoder Layers	12	32
Attention Heads	12	18
Dropout	0.1	0.1
General Hyperparameters		
Batch Size	32	32
Epochs	4 (Curriculum)	4 (Curriculum)
AdamW ϵ	1×10^{-5}	1×10^{-5}
Weight Decay	0.01	0.01
Warm-up Ratio	0.1	0.1
Gradient Clipping	1.0	1.0
Differential Learning Rates		
Embedding + First 3 Layers	1×10^{-6}	–
Remaining Encoder Layers	1×10^{-5}	1×10^{-5}
Self-Attention + Head	5×10^{-5}	5×10^{-5}

Table 3: Model architecture and training hyperparameters for both encoder backbones.

B Prompts

To illustrate our methodology, we provide representative 1-shot examples of the prompt templates used for both subtasks. While our experimental setup utilized 7-shot prompting, these examples highlight the structural integration of the text and its corresponding phonetic (IPA) transcription.

B.1 Subtask 1: Polarization Detection

System: You are a strict classifier. Categorize whether the input text is polarized or not.

User:
Text: Hawa Ni moja ya Watu wabaya sana Mufirisi tunaowaabudu hapa Tz WAHINDI WAARABU WACHINA Ameniambia Mzee Hussen hapa Ila AnyWay
IPA: hawa ni moa ja watu waaja sana mufiisi tunaowaauu hapa tz wahindi waaau watina ameniambia mzee hussen hapa ila awaj
Is the text Polarized?

Assistant: YES

User:
Text: hao ni wakikuyu wanyonge sana sisi wengine hatukupigia kura huko
IPA: hao ni wakikuju waone sana sisi wenine hatukupiia kua huko
Is the text Polarized?

Assistant:

B.2 Subtask 2: Polarization Type Detection

System: You are a strict classifier. Identify the categories of polarization present in the text.

User:
Text: Hawa Ni moja ya Watu wabaya sana Mufirisi tunaowaabudu hapa Tz WAHINDI WAARABU WACHINA Ameniambia Mzee Hussen hapa Ila AnyWay
IPA: hawa ni moa ja watu waaja sana mufiisi tunaowaauu hapa tz wahindi waaau watina ameniambia mzee hussen hapa ila awaj

Assistant:
Labels (political, racial, religious, gender, other): NO, YES, NO, NO, NO

User:
Text: hao ni wakikuyu wanyonge sana sisi wengine hatukupigia kura huko
IPA: hao ni wakikuju waone sana sisi wenine hatukupiia kua huko

Assistant:
Labels (political, racial, religious, gender, other):