SMASH at SemEval-2026 Task 9: Detecting Multilingual Polarisation with Encoder Ensembles and Calibrated Decision Thresholds

Zahra Bokaei*, **Alessandra Terranova***, **Yi Zheng***, **Tom Bidewell, Björn Ross**
School of Informatics, University of Edinburgh
zahra.bokaei, a.terranova@ed.ac.uk; Y.Zheng-77, T.Bidewell@sms.ed.ac.uk; b.ross@ed.ac.uk

Abstract

This paper describes the SMASH submission to SemEval-2026 Task 9 on multilingual, multicultural, and multi-event polarisation detection. The task comprises (i) binary polarisation detection, (ii) multi-label classification of polarisation types, and (iii) multi-label identification of polarisation manifestations across all available languages. We propose a language-adaptive ensemble framework combining monolingual and multilingual encoder-only transformers, together with a principled out-of-fold (OOF) threshold tuning strategy. Instead of relying on fixed probability thresholds, we jointly tune ensemble weights and class-wise decision thresholds to directly optimise macro-F1 under the official evaluation metric. Our experiments show that (1) monolingual encoders dominate in several high-resource languages but benefit from complementary multilingual signals, (2) no single multilingual backbone universally outperforms others across languages and subtasks, and (3) language-specific class threshold tuning substantially improves performance due to large cross-lingual variation in class distributions. Our results demonstrate that careful logit-level ensembling and threshold tuning provide strong performance for multilingual, imbalanced, multi-label polarisation detection. Across 22 evaluation languages, SMASH ranks among the top three systems in a substantial number of language–subtask pairs. Specifically, it ranks in the top three for 5 languages in Subtask 1, 14 languages in Subtask 2, and 16 languages in Subtask 3, demonstrating strong and consistent performance across diverse languages and tasks. Our system achieves average macro-F1 scores of 0.81, 0.62, and 0.53 for Subtasks 1, 2, and 3, respectively.

1 Introduction

Online polarisation has become a defining characteristic of contemporary social media discourse, where opinions increasingly manifest as sharp divisions between opposing groups. Polarised content is not simply strongly opinionated language; rather, it reflects attitude polarisation, characterised by hostility toward out-groups, blind support for in-groups, stereotyping, dehumanization, and intolerance (Lerman et al., 2024). Accurately detecting such phenomena across languages and cultural contexts remains a challenging problem for NLP systems, particularly when polarisation is expressed implicitly or encoded in cultural references.

SemEval-2026 Task 9 (Naseem et al., 2026a,b) addresses this challenge by introducing a multilingual, multicultural, and multi-event benchmark for polarisation detection across 22 languages. The task comprises three subtasks: (1) binary polarisation detection, (2) multi-label classification of polarisation types, and (3) multi-label identification of polarisation manifestations. The multi-label nature of Subtasks 2 and 3, combined with substantial cross-lingual variation and class imbalance, makes macro-F1 optimisation particularly difficult. Moreover, some languages benefit from strong monolingual pretrained models, while others rely primarily on multilingual representations.

In this work[1], we investigate how careful logit-level ensembling and language-specific decision threshold tuning can improve multilingual polarisation detection without resorting to large instruction-tuned generative models. We fine-tune multiple encoder-only transformers, including three multilingual backbones (mDeBERTa-v3 (He et al., 2021), XLM-RoBERTa-Large (Conneau et al., 2019), mmBERT (Marone et al., 2025)) and language-specific monolingual encoders for high-resource languages. Instead of relying on a fixed probability threshold, we perform 5-fold cross-validation to obtain out-of-fold (OOF) logits and

*These authors contributed equally to this work.

[1] https://github.com/WendyZheng9953/SMASH-SemEval-2026-Task-9

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 832–848

jointly tune ensemble weights and per-label decision thresholds to directly optimise macro-F1 for each language and subtask.

Our experiments show that: (1) monolingual encoders dominate in several high-resource settings but benefit from multilingual complementary signals; (2) the optimal multilingual backbone varies by language and subtask, indicating that no single model universally dominates; and (3) language-specific threshold tuning yields substantial improvements, as optimal thresholds vary dramatically across languages and labels. These findings highlight that ensemble design and explicit per-label decision threshold tuning can be critical for multilingual, imbalanced, multi-label classification tasks. In summary, our contributions are: (i) a language-adaptive ensemble framework combining monolingual and multilingual encoders; (ii) a principled OOF-based strategy for jointly tuning ensemble weights and class thresholds under the official macro-F1 metric; (iii) an empirical analysis demonstrating the importance of language-specific tuning across 22 languages and three subtasks; and (iv) a careful error analysis of selected languages.

2 Background

2.1 Polarisation and the POLAR Task

Polarisation on social media refers to the process in which opinions or behaviours become more extreme and divided, leading to a greater conflict or distance between differing groups (Sunstein, 2018). Online polarisation can threaten social cohesion, aided by the role of social media as echo chambers and by the presence of biased content (Waller and Anderson, 2021). Importantly, a better understanding of polarisation could help minimise its negative impact on mainstream politics, democratic decision making, and society (Ali et al., 2025).

Polarisation is embedded in complex social narratives and cultural specificity, but most work in NLP has focused on English or other high-resource languages, specific domains, or binary classification. The POLAR@SemEval-2026 shared task addresses these gaps and the need for automated systems capable of characterizing polarisation across different linguistic and cultural contexts (Naseem et al., 2026a,b). The authors decompose the phenomenon into three distinct sub-problems:

1. Subtask 1, Detection: Binary classification determining if a text has attitude polarisation.
2. Subtask 2, Type Classification: Multi-label classification identifying the target dimension, such as political or racial/ethnic.
3. Subtask 3, Manifestation Identification: Multi-label classification identifying rhetorical strategies like dehumanization or vilification.

Performance is evaluated using macro-F1 across 22 languages, ranging from high-resource like English or Spanish, to low-resource such as Amharic, Hausa, or Odia.

2.2 Multilingual Encoder Models

The detection of nuanced social phenomena like polarisation across multiple languages necessitates models capable of capturing both universal semantic structures and language-specific expressions. Early multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) established that large-scale pretraining enables robust cross-lingual transfer. The modern landscape has further evolved with models like mmBERT(Marone et al., 2025), which employs annealed language learning and variable masking to bridge the gap for low-resource languages.

However, these models may face the "curse of multilinguality" (Pfeiffer et al., 2022) and language interference, with performance of high-resource languages lagging behind highly-resourced monolingual models. Monolingual models still hold a significant advantage in high-resource settings by capturing specialised jargon and cultural references without the interference of a shared multilingual feature space (Nozza et al., 2020). Recent work in Persian toxic language detection (Bokaei et al., 2025, 2026) also indicates that cross-lingual transfer gains are sensitive to cultural proximity between languages (Bokaei et al., 2025), reinforcing the importance of language-aware modeling strategies.

2.3 Model Ensemble and Threshold Tuning

Model ensembling is a critical technique for improving the robustness and performance of NLP systems (Jia et al., 2023). By aggregating the predictions of multiple models, for example aiming to combine the linguistic depth of specialised monolingual models with the broad semantic understanding of large multilingual models, ensembles can effectively mitigate the errors of individual architectures.

Model ensembling is particularly successful in competitive shared tasks like SemEval (Abaskohi

et al., 2024; Li et al., 2024; Le et al., 2025). Our approach distinguishes itself by implementing logit-level ensembling followed by metric-aware threshold calibration. In multi-label and imbalanced classification settings typical of the POLAR dataset, simple probability averaging often fails to capture minority classes. By optimizing thresholds specifically for the Macro-F1 on a held-out development set, we ensure our system remains sensitive to rare but critical manifestations of polarised discourse.

2.4 POLAR Dataset Analysis

The POLAR dataset exhibits substantial class imbalance across all 22 languages. Subtask 1 prevalence ranges from 10.7% (Hausa) to 90.8% (Khmer); Subtask 2 shows extreme within-language skew (e.g., Amharic Political 66.8% vs. Gender/Sexual 0.6%); and Subtask 3 manifestation patterns vary by culture, with Dehumanization rarest globally (median $<8\%$). These patterns motivate class weighting and language-specific calibration. Full statistics are in Appendix A.

3 System Overview

3.1 Base Models

We fine-tune both monolingual and multilingual encoders (mDeBERTa, XLM-R, and mmBERT) separately for all three subtasks. For high-resource languages, we additionally fine-tune a monolingual encoder on the language-specific training data (Table 5). The monolingual model selection was based on their classification performance on the dev set. Each subtask is trained with a prediction head: a binary classifier for Subtask 1 and independent sigmoid outputs for Subtasks 2–3. Our approach does not introduce additional computational overhead beyond standard fine-tuning. Fine-tuning details provided in Appendix C.

3.2 Cross-Validation and Out-of-Fold Logits

We perform K-fold cross-validation ($K = 5$) on the official training data for each language and subtask to select ensemble weights. In the development phase, we use the official development data for testing. For each fold, we fine-tune models with three random seeds and average the resulting validation logits to reduce variance. We then concatenate all folds to obtain out-of-fold (OOF) logits covering the full training set. These OOF logits form the basis for tuning both model ensemble weights and class decision thresholds.

3.3 Model Ensembling

First, we apply language-specific logit-level ensembling and reduce variance by averaging across three random seeds for each model. For each backbone model M (mono- or multilingual) fine-tuned with seeds $\mathcal{S} = \{1, 42, 111\}$, we average their logits:

$$\bar{\mathbf{z}}_i^{(M)} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{z}_i^{(M,s)}. \quad (1)$$

Then, for low-resource languages, we ensemble the three multilingual models using weights that sum to one:

$$\bar{\mathbf{z}}_i^{(\text{ens})}(\mathbf{w}) = \sum_{b=1}^{3} w_b \, \bar{\mathbf{z}}_i^{(M_b)}, \; \sum_{b=1}^{3} w_b = 1, \quad (2)$$

where $\bar{\mathbf{z}}_i^{(M_b)}$ are the logits averaged across seeds from multilingual model M_b. Meanwhile, for high-resource languages, we ensemble a monolingual encoder with one multilingual encoder using a single weight $w \in [0, 1]$:

$$\bar{\mathbf{z}}_i^{(\text{ens})}(w) = w \, \bar{\mathbf{z}}_i^{(\text{mono})} + (1 - w) \, \bar{\mathbf{z}}_i^{(\text{multi})}. \quad (3)$$

3.4 Tuning Ensemble Weights and Thresholds

We tune ensemble weights and class decision thresholds jointly on out-of-fold (OOF) predictions to directly optimise macro-F1. The OOF threshold calibration procedure ensures that thresholds are optimised on held-out data rather than training labels, mitigating overfitting. For each language and fold, we obtain validation logits for every model and seed, compute seed-averaged logits $\bar{\mathbf{z}}^{(M)}$, and then concatenate folds to form merged OOF logits. We grid-search candidate ensemble weights (optimisation step of 0.05). For each candidate weight setting, we compute ensembled OOF logits $\bar{\mathbf{z}}^{(\text{ens})}$ and select class thresholds that maximise macro-F1 on the merged OOF set. For all subtasks we choose class-wise thresholds $\mathbf{t} = (t_1, \ldots, t_L)$, with optimising step of 0.01:

$$t_\ell = \arg\max_{t \in [0,1]} F1_\ell(t), \qquad \ell = 1, \ldots, L. \quad (4)$$

3.5 Final Fine-tuning and Inference

For the final submission, we fine-tune both the mono- and multilingual models on the full training data using three random seeds (same as development), average seed logits at inference, apply the learnt language-specific model ensembling weights, and then apply the tuned class threshold(s) to produce the final predictions. This design combines the strengths of language-specialised representations and multilingual generalisation.

4 Results

We compare our system with the official leaderboard and the baseline systems provided by the shared task organisers. The baselines are fine-tuned LaBSE (Feng et al., 2022) using the training data for each language for all subtasks. Table 1 reports macro-F1 scores across all languages and subtasks for the Baseline, the top-ranked system (1st Rank), and our model (SMASH). Overall, SMASH achieves competitive performance, ranking among the top systems across multiple languages and subtasks, and obtaining the best results in several cases (highlighted in bold). It consistently outperforms the baseline across all subtasks. Moreover, SMASH ranks within the top three systems in a substantial number of language–subtask pairs: 5 languages in Task 1, 14 in Task 2, and 16.

4.1 Analysis of Performance

Model ensemble weight analysis. The optimal ensemble composition varies substantially by language and subtask. For high-resource languages, the mono+multi ensemble typically assigns the majority weight to the monolingual encoder ($w_{\text{mono}} \in [0.55, 0.90]$), while the selected multilingual model depends on the language: XLM-R is chosen for most cases, with mmBERT selected for English in Subtasks 1–2, and mDeBERTa selected for Arabic and Chinese. This pattern suggests that monolingual models capture strong in-language signals, while the multilingual model provides complementary generalisation that is most beneficial for particular languages. For low-resource languages, the XLM-R frequently has the largest weight; however, mDeBERTa becomes dominant for several languages and subtasks (e.g., Subtask 1 and Subtask 3 for swa), indicating that the optimal backbone is highly language-dependent. Appendix E report the ensemble weights and class thresholds.

Class threshold analysis. The tuned decision thresholds exhibit large cross-lingual variation, highlighting the importance of language-specific threshold tuning. In Subtask 1, the optimal threshold ranges from 0.02 to 0.99, implying substantial differences in class priors across languages. For Subtasks 2–3, class-wise thresholds also vary widely. This indicates that a fixed threshold of 0.5 would not be optimal for many languages and classes, motivating our tuning of model ensemble weights and thresholds on OOF logits.

4.2 Error Analysis

In this section, we focus on two languages: English, as a high-resource language, and Persian (Farsi), which achieved first place in Subtasks 2 and 3. We conduct an error analysis for both settings (confusion matrix shown in Appendix F). Error analysis helps us understand where the models succeed and where they struggle. We manually reviewed both correctly predicted and misclassified examples across the three subtasks, paying particular attention to recurring patterns, label-specific difficulties, and common sources of confusion.

4.2.1 English Error Analysing

Cases detected. For Subtask 1, detected posts are often overtly polarised, featuring strong negative sentiment (e.g., slurs) and explicit out-group labelling (e.g., "nazis", "fascist"). For Subtask 2, the detected type labels usually appear when the target is explicit. For example, directly naming the political ideologies and religions (e.g., "far right", "Christian"). For Subtask 3, the detected manifestations are expressed in a straightforward way. For example, explicit insults or clear group generalisations. *Dehumanization* and *lack_of_empathy* are detected when the tone is clearly indifferent and often calling for extreme actions or expressing alarmist rhetoric ("deport to x", "ruin x").

Cases missed. For all Subtasks, most errors reflect label ambiguity as the classifiers only predict on the text without enough context—cases where reasonable annotators might get confused. More conversation context could help resolve this ambiguity and reduce noisy labels. In Subtask 1, false negatives tend to use more implicit language and depend on context, often expressed via sarcastic rhetorical questions, or a discussion-like tone where polarisation is conveyed through implication (e.g., immigration framing). In Subtask 2, *religious* and *gender* are rarest labels in train data thus lowest F1. Missed cases are often expressed indirectly or blended into broader ideology ("blue/red states"). These missed cases require knowledge about historical and current events to guess the writer's implications. In Subtask 3, *lack_of_empathy* has the lowest F1. Missed cases are subtle (phrased as "reasonable-sounding" commentary or correcting someone's "misbelief"), which likely requires pragmatic understanding.

Subtask	System	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
1	Baseline	0.715	0.795	0.852	0.671	0.780	0.842	0.775	0.737	0.677	0.659	0.821
	1st Rank	0.800	0.848	0.862	0.760	0.825	0.834	0.833	0.823	0.730	0.774	0.891
	SMASH	0.775	0.828	0.831	0.720	0.797	0.821	0.811	0.811	0.605	**0.774**	0.885
2	Baseline	0.371	0.485	0.288	0.407	0.333	0.462	0.203	0.791	0.375	0.626	0.477
	1st Rank	0.669	0.669	0.421	0.620	0.532	0.643	0.479	0.807	0.550	0.704	0.747
	SMASH	0.650	0.658	0.378	0.575	0.487	**0.643**	0.454	**0.807**	0.467	0.702	0.736
3	Baseline	0.443	0.390	0.086	0.348	0.410	0.200	0.745	0.234	–	0.609	–
	1st Rank	0.579	0.645	0.280	0.517	0.510	0.493	0.207	0.770	–	0.437	–
	SMASH	**0.579**	0.641	**0.280**	0.513	0.507	**0.493**	0.201	**0.770**	–	**0.437**	–
Subtask	System	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
1	Baseline	0.879	0.776	0.789	0.724	0.745	0.726	0.757	0.644	0.695	0.789	0.869
	1st Rank	0.923	0.825	0.825	0.843	0.830	0.803	0.811	0.905	0.832	0.819	0.931
	SMASH	0.914	0.808	0.799	0.828	0.808	0.798	0.810	0.901	0.787	0.795	0.903
2	Baseline	0.721	0.560	0.365	0.449	0.645	0.593	0.441	0.314	0.470	0.712	0.669
	1st Rank	0.810	0.602	0.552	0.649	0.590	0.680	0.569	0.573	0.776	0.797	0.843
	SMASH	0.779	0.577	**0.552**	0.617	0.619	0.673	**0.569**	0.458	0.619	0.790	0.817
3	Baseline	0.131	0.384	0.456	–	–	0.508	0.220	0.673	0.769	0.531	0.000
	1st Rank	0.712	0.329	0.544	–	–	0.540	0.584	0.444	0.538	0.821	0.719
	SMASH	0.711	**0.329**	0.541	–	–	**0.540**	**0.584**	**0.444**	0.514	0.821	0.677

Table 1: Test set macro-F1 of all languages and subtasks, with Baseline, 1st Rank, and SMASH (ours). Scores where SMASH ranks 1st are shown in bold.

4.2.2 Persian Error Analysing

Cases detected. For Subtask 1, detected Persian posts are typically overtly polarised, featuring explicit evaluative language, profanity, or clear in-group/out-group framing (e.g., اخوند وطن فروش, cleric, traitor) Successful detections often involve direct hostility toward political actors or institutions, especially when combined with strong affective markers or explicit endorsement of punishment (e.g., اعدام باید گردد -must be executed). For Subtask 2, rare labels in the training data (*race*, *gender*) are correctly detected when expressed explicitly and with salient lexical cues. Race-related posts include overt national or ethnic markers, while gender cases explicitly reference women's identity. For Subtask 3, *lack of empathy* is the rarest label. This class is correctly detected when posts contain explicit calls for violence or punishment (e.g., اعدام آتش بزن "Execution", "set on fire"), dehumanizing language, or clear endorsement of retaliatory harm. Detection is strongest when harm is advocated directly and lexically unambiguous.

Cases missed. In Subtask 1, false negatives frequently rely on implicit polarisation expressed through report-like framing, abstract political discussion, or sarcastic commentary. Rather than explicit insults, polarisation is conveyed via contextual topic selection (e.g., references to زندان اوین اعتصاب ها فیلترینگ) Evin prison, strikes, filtering or through irony and intertextual cultural cues, which require pragmatic interpretation. In Subtask 2, *race* and *gender* perform more weakly due to their rarity. Missed cases often imply identity through broader ideological or institutional critique (e.g., references to nationality, migration, or cultural norms) without explicit naming, requiring sociopolitical context to interpret. In Subtask 3, missed *lack of empathy* instances frequently normalise or trivialise harm, often framed as descriptive commentary, sarcasm, or exaggeration. The absence of explicit violence cues makes them harder to detect.

The error patterns observed across English and Persian suggest several concrete directions for improving polarisation detection. First, since many false negatives in Subtask 1 rely on implicit framing, sarcasm, or rhetorical questions, incorporating broader conversational or thread-level context (e.g., parent posts, reply chains) could help disambiguate these cases. Second, the persistent under-detection of rare labels in Subtask 2 (*religious*, *gender*, *race*) motivates targeted data augmentation or class-balanced sampling to strengthen minority-class signals. Third, errors involving cultural or sociopolitical references (e.g., Persian references to Evin prison, English references to "blue/red states") indicate that integrating external knowledge through entity linking or retrieval-augmented inputs could improve pragmatic interpretation. Finally, the difficulty of detecting *lack of empathy* suggests that fine-grained pragmatic features or LLM-based rationalisation could help capture indirect harm endorsement. Representative examples of correctly classified and missed predictions are

provided in Appendix K.

4.3 PCA Analysis of Sentence Embeddings

We conduct principal component analysis to better understand cross-lingual structure. The results are shown in Appendix G.

Comparison across models. The PCA projections show different multilingual geometries across the three models before fine-tuning. mDeBERTa exhibits the strongest global mixing: languages densely overlap, suggesting a language-agnostic shared space. XLM-R produces a broad embedding space where languages overlap heavily, but some are still denser in particular regions. In contrast, mmBERT displays a language-stratified structure, with prominent clusters and long-tail patterns dominated by a subset of languages, indicating that language identity remains comparatively salient in its embedding space. These differences are consistent with the behaviour we observe in model ensembling: XLM-R and mDeBERTa often provide robust cross-lingual generalisation.

Before vs. after fine-tuning on Subtask 2. Comparing each model before and after fine-tuning in Subtask 2 shows that fine-tuning reshapes the embedding spaces in a systematic way. After fine-tuning, mDeBERTa and XLM-R become more globally spread and intermixed. mmBERT also changes under fine-tuning but retains clearer language-specific clustering, implying that fine-tuning does not fully wash out language-identity signals. Taken together, results suggest that task-specific fine-tuning improves cross-lingual alignment in the embedding space, but the extent of this alignment is model-dependent. This motivates our use of language-aware model ensembling rather than a single model choice.

5 Conclusion

We presented SMASH, our submission to SemEval-2026 Task 9 on multilingual polarisation detection. Our results demonstrate that careful class threshold calibration and language-aware ensembling can substantially improve performance in multilingual, imbalanced, and multi-label settings. A key insight from our experiments is the contrasting behaviour between high- and lower-resource languages. In high-resource settings, monolingual encoders consistently capture strong in-language signals and often dominate ensemble composition, while multilingual models provide complementary generalisation. In contrast, lower-resource languages rely more heavily on multilingual backbones, although the optimal architecture varies by language and subtask. These findings suggest that multilingual performance is not only determined by architecture but by how well the representations align with language-specific discourse patterns. The extreme variation in optimal decision thresholds across languages highlights the importance of metric-aware threshold tuning. A fixed threshold is often suboptimal, particularly in macro-F1 evaluation, reinforcing the need for language-specific tuning strategies.

Limitations

In this work, we use the terms "high" and "low-resource" in a task-specific sense rather than as a statement about the global availability of NLP resources for a language. We probed monolingual encoders on the dev set and classified a language as high-resource for our system if a monolingual model consistently outperformed the multilingual models. Otherwise, we treated it as low-resource. As a result, this categorisation depends on the particular set of mono-lingual models we evaluated, it does not necessarily correlate with dataset size. For example, German is commonly considered as high-resource, but in experiments we did not identify a mono-lingual model that reliably surpassed the multilingual baselines, and we therefore handled German with the low-resource strategy. This highlights a limitation of our approach: our high/low-resource split reflects model availability and empirical performance under our pipeline.

Ethical Considerations

Polarisation is inherently context-dependent and the provided annotations reflect subjective human judgments, which may vary across annotators, social communities and cultures. As a result, models trained on these annotations may produce biases and should not be interpreted as measuring a universal notion of polarisation. Moreover, automatic predictions can be sensitive to sarcasm or implicit context, and may yield false positives or false negatives. It could be harmful if these models were misused for content moderation. Therefore, our system is not intended for fully automated polarisation classification. Instead, it should be used as a support tool to assist human analysis with consideration of potential downstream impacts.

Acknowledgements

This research was supported by the UKRI AI Centre for Doctoral Training in Responsible and Trustworthy in-the-world Natural Language Processing (grant ref: EP/Y030656/1).

References

Amirhossein Abaskohi, Amirhossein Dabiriaghdam, Lele Wang, and Giuseppe Carenini. 2024. BCAmirs at SemEval-2024 task 4: Beyond words: A multimodal and multilingual exploration of persuasion in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1412–1423, Mexico City, Mexico. Association for Computational Linguistics.

Adem Chanie Ali, Seid Muhie Yimam, Abinew Ali Ayele, Chris Biemann, and Martin Semmann. 2025. Silenced voices: social media polarization and women's marginalization in peacebuilding during the northern ethiopia war. *i-com*, 24(2):407–432.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9290–9304, Vienna, Austria. Association for Computational Linguistics.

Zahra Bokaei, Bonnie Webber, and Walid Magdy. 2026. Benchmarking offensive language detection in Persian and Pashto. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 13–23, Rabat, Morocco. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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

Jianguo Jia, Wen Liang, and Youzhi Liang. 2023. A review of hybrid and ensemble in deep learning for natural language processing. *arXiv preprint arXiv:2312.05589*.

Tung Thanh Le, Tri Minh Ngo, and Trung Hieu Dang. 2025. Anastasia at SemEval-2025 task 9: Subtask 1, ensemble learning with data augmentation and focal loss for food risk classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 141–147, Vienna, Austria. Association for Computational Linguistics.

Kristina Lerman, Dan Feldman, Zihao He, and Ashwin Rao. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(1):8.

Dailin Li, Chuhan Wang, Xin Zou, Junlong Wang, Peng Chen, Jian Wang, Liang Yang, and Hongfei Lin. 2024. CoT-based data augmentation strategy for persuasion techniques detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1315–1321, Mexico City, Mexico. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. 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, "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. *arXiv preprint arXiv:2505.20624*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495.

Cass R Sunstein. 2018. Republic: Divided democracy in the age of social media.

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian. *Preprint*, arXiv:2309.10931.

A POLAR Dataset Analysis

Tables 2, 3, and 4 present the class statistics for each language across all three subtasks. This appendix expands on the dataset analysis summarised in Section 2.4.

Subtask 1 (Binary polarisation detection). The training data exhibits substantial class imbalance across the 22 languages. Several languages demonstrate extreme positive skew, where the majority of samples are labelled as polarised: Khmer (90.8%), Hindi (85.5%), Amharic (75.6%), and Persian (74.1%). Conversely, Hausa (10.7%) and Odia (28.8%) show extreme negative skew, where polarised content is rare and models may learn to systematically ignore positive instances. The severe imbalance in high-skew languages motivates the use of class weighting, focal loss, or resampling strategies during training.

Subtask 2 (Polarisation type classification). The data reveals pronounced imbalance both within and across languages. For example, Amharic exhibits extreme imbalance, with Political polarisation dominating at 66.8% while Gender/Sexual polarisation appears in only 0.6% of samples. Applying uniform classification thresholds across all labels will systematically miss these rare categories, suggesting the need for per-label threshold optimisation.

Subtask 3 (Manifestation classification). The data reveals distinct patterns of polarisation expression that vary across linguistic contexts. Urdu demonstrates uniformly high prevalence across all six manifestations, while Hausa exhibits extreme sparsity with all manifestations below 11%—a consequence of the class imbalance observed in Subtask 1. Dehumanization emerges as the rarest manifestation globally, with median prevalence below 8% across languages, making it particularly difficult to detect. Persian data exhibits more Vilification manifestations (57.5%) with relatively low Stereotype (13.1%), while Khmer shows the inverse pattern. These statistics indicate that polarisation manifestation expressions vary across culture, and models trained on one language may not transfer effectively to others. This further suggests the need for language-specific calibration to achieve robust classification performance across languages.

Lang	Total	% Not Polar	% Polar
amh	3,332	24.43	75.57
arb	3,380	55.27	44.73
ben	3,333	57.28	42.72
deu	3,180	52.45	47.55
eng	3,222	63.53	36.47
fas	3,295	25.95	74.05
hau	3,651	89.26	10.74
hin	2,744	14.50	85.50
ita	3,334	58.97	41.03
khm	6,640	9.20	90.80
mya	2,889	41.78	58.22
nep	2,005	49.73	50.27
ori	2,368	71.16	28.84
pan	1,700	50.59	49.41
pol	2,391	58.05	41.95
rus	3,348	69.44	30.56
spa	3,305	49.77	50.23
swa	6,991	49.88	50.12
tel	2,366	46.15	53.85
tur	2,364	51.14	48.86
urd	3,563	30.51	69.49
zho	4,280	50.44	49.56
Total	66,378	45.97	54.03

Table 2: Training data statistics for Subtask 1.

Language	Total	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
amh	3,332	66.81	25.93	1.98	0.57	24.79
arb	3,380	23.08	17.25	8.37	10.92	16.72
ben	3,333	34.02	0.75	1.95	0.51	10.05
deu	3,180	40.72	18.52	11.13	5.88	13.81
eng	3,222	35.69	8.72	3.48	2.23	3.91
fas	3,295	43.92	2.43	9.62	5.98	24.22
hau	3,651	4.88	3.15	2.55	0.79	0.38
hin	2,744	73.72	12.14	58.71	11.48	13.12
ita	3,334	0.00	22.38	8.55	11.43	0.00
khm	6,640	18.33	1.48	3.37	1.70	65.92
mya	2,889	25.27	5.26	3.08	10.59	45.14
nep	2,005	17.21	14.01	7.93	5.24	11.77
ori	2,368	20.95	5.03	6.33	3.34	3.67
pan	1,700	30.76	5.88	7.88	11.18	8.94
pol	2,391	36.60	8.99	3.64	4.60	6.48
rus	3,348	13.86	9.83	4.09	5.65	2.36
spa	3,305	27.26	18.85	15.89	13.40	13.40
swa	6,991	2.66	35.52	3.53	2.23	7.94
tel	2,366	21.60	16.99	8.96	13.27	23.71
tur	2,364	44.71	16.92	15.23	4.78	4.82
urd	3,563	67.22	54.39	55.26	51.22	50.74
zho	4,280	5.86	22.62	1.99	16.89	8.60

Table 3: Training data statistics for Subtask 2.

Language	Total	Stereotype	Vilification	Dehuman.	Extreme Lang.	Lack Empathy	Invalidation
amh	3,332	54.56	48.50	13.15	30.55	17.59	16.00
arb	3,380	33.34	37.16	10.95	30.38	17.01	8.11
ben	3,333	5.97	24.06	10.71	4.71	1.89	1.77
deu	3,180	35.85	30.06	14.94	21.76	26.67	16.26
eng	3,222	15.11	26.63	12.14	23.90	11.08	18.19
fas	3,295	13.14	57.48	4.31	16.90	9.86	7.98
hau	3,651	4.27	1.23	3.48	3.01	0.88	0.25
hin	2,744	49.71	65.16	18.22	50.58	56.74	65.67
khm	6,640	68.28	1.52	1.22	2.26	10.98	6.54
nep	2,005	26.78	31.42	6.58	27.13	10.57	14.96
ori	2,368	9.97	11.70	0.68	13.39	1.56	3.38
pan	1,700	16.24	40.41	22.00	23.94	12.41	24.41
spa	3,305	27.47	30.59	8.93	24.18	23.93	10.62
swa	6,991	39.69	41.24	12.77	23.93	29.75	23.42
tel	2,366	11.20	22.65	2.49	13.44	26.29	22.78
tur	2,364	40.82	32.36	10.87	43.10	9.56	4.02
urd	3,563	62.25	64.75	55.63	62.17	56.24	57.23
zho	4,280	30.07	18.50	5.02	8.13	7.87	4.79

Table 4: Manifestation Identification Task - Training Data Statistics (%)

B Monolingual models for high-resource languages

Table 5 lists the monolingual models chosen for high-resource languages.

Lang	Model
arb	bert-large-arabertv02-twitter (Antoun et al.)
eng	bertweet-large (Nguyen et al., 2020)
pol	herbert-large-cased (Mroczkowski et al., 2021)
rus	ruRoberta-large (Zmitrovich et al., 2023)
spa	robertuito-base-uncased (Pérez et al., 2022)
zho	chinese-roberta-wwm-ext-large (Cui et al., 2020)

Table 5: Monolingual models for high-resource languages.

C Fine-tuning details

All models for Subtask 1, 2 and 3 were fine-tuned under this hyperparameter setting: max input length of 128, epochs of 10, learning rate of 1e-5 and train batch size of 16. We append a short language encoding text "[lang=amh/arb/...]" before each text instance in fine-tuning and inference.

D Our system's ranking on the official leaderboard

Our system's ranking on the official leaderboard is shown in Table 6. We rank top 10 in half of the languages for Subtask 1. We rank top 5 in most languages for Subtask 2 and 3.

Task	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
1	9	13	19	12	17	5	8	11	18	1	4
2	3	3	3	5	9	1	2	1	5	2	2
3	1	2	1	3	3	1	4	1	–	1	–
Task	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
1	3	6	5	4	4	3	2	2	12	5	9
2	5	4	1	4	3	3	1	3	5	2	7
3	2	1	2	–	–	1	1	1	4	2	3

Table 6: Our system's ranks of all languages and subtasks.

E Ensemble weights and class decision thresholds

Table 7 and 8 report the model ensemble weights and class decision thresholds after grid search.

Task	Lang	Multi	w_mono	w_multi	t_pol					
1	arb	mdeberta	0.50	0.50	0.16					
	eng	mmbert	0.80	0.20	0.72					
	pol	xlm	0.55	0.45	0.42					
	rus	xlm	0.70	0.30	0.48					
	spa	xlm	0.65	0.35	0.46					
	zho	xlm	0.50	0.50	0.74					
Task	Lang	Multi	w_mono	w_multi	t_pol	t_rac	t_rel	t_gen	t_oth	
2	arb	xlm	0.75	0.25	0.43	0.44	0.47	0.39	0.38	
	eng	mmbert	0.70	0.30	0.27	0.23	0.28	0.16	0.27	
	pol	xlm	0.70	0.30	0.42	0.22	0.47	0.39	0.39	
	rus	xlm	0.60	0.40	0.61	0.31	0.37	0.4	0.34	
	spa	xlm	0.60	0.40	0.49	0.46	0.46	0.44	0.46	
	zho	mdeberta	0.70	0.30	0.34	0.56	0.36	0.48	0.34	
Task	Lang	Multi	w_mono	w_multi	t_ste	t_vil	t_deh	t_ext	t_lac	t_inv
3	arb	mdeberta	0.80	0.20	0.46	0.46	0.36	0.5	0.4	0.47
	eng	xlm	0.90	0.10	0.52	0.51	0.48	0.47	0.44	0.46
	spa	xlm	0.60	0.40	0.48	0.45	0.37	0.47	0.4	0.49
	zho	xlm	0.40	0.60	0.48	0.43	0.31	0.45	0.42	0.38

Table 7: Model ensemble weights ("w_") and class decision thresholds ("t_") for all high-resource languages across three subtasks. Label abbreviations: pol=Political, rac=Racial, rel=Religious, gen=Gender, oth=Other; ste=Stereotype, vil=Vilification, deh=Dehumanization, ext=Extreme Language, lac=Lack of Empathy, inv=Invalidation.

Task	Lang	w_xlm	w_mmbert	w_mdeberta	t_pol					
1	amh	0.50	0.20	0.30	0.78					
	ben	0.90	0.00	0.10	0.13					
	deu	0.50	0.10	0.40	0.07					
	fas	0.40	0.20	0.40	0.93					
	hau	0.30	0.20	0.50	0.02					
	hin	0.50	0.10	0.40	0.76					
	ita	0.60	0.10	0.30	0.26					
	khm	0.80	0.00	0.20	0.99					
	mya	0.20	0.20	0.60	0.60					
	nep	0.60	0.10	0.30	0.71					
	ori	1.00	0.00	0.00	0.03					
	pan	0.50	0.10	0.40	0.31					
	swa	0.20	0.20	0.60	0.80					
	tel	0.50	0.10	0.40	0.90					
	tur	0.50	0.20	0.30	0.66					
	urd	0.70	0.10	0.20	0.30					
Task	Lang	w_xlm	w_mmbert	w_mdeberta	t_pol	t_rac	t_rel	t_gen	t_oth	
2	amh	0.40	0.00	0.60	0.40	0.49	0.41	0.32	0.48	
	ben	1.00	0.00	0.00	0.36	0.20	0.52	0.38	0.47	
	deu	0.40	0.10	0.50	0.24	0.34	0.43	0.23	0.30	
	fas	0.60	0.10	0.30	0.41	0.32	0.59	0.44	0.54	
	hau	0.60	0.20	0.20	0.31	0.19	0.23	0.08	0.08	
	hin	0.50	0.10	0.40	0.56	0.44	0.6	0.51	0.35	
	ita	0.50	0.10	0.40	0.01	0.34	0.35	0.54	0.01	
	khm	0.50	0.00	0.50	0.49	0.40	0.42	0.31	0.61	
	mya	0.60	0.10	0.30	0.54	0.33	0.34	0.27	0.60	
	nep	0.60	0.10	0.30	0.29	0.73	0.27	0.71	0.31	
	ori	0.80	0.10	0.10	0.17	0.17	0.25	0.24	0.25	
	pan	0.60	0.10	0.30	0.57	0.24	0.33	0.47	0.22	
	swa	0.60	0.00	0.40	0.42	0.66	0.74	0.36	0.39	
	tel	0.60	0.10	0.30	0.36	0.42	0.41	0.30	0.38	
	tur	0.80	0.20	0.00	0.31	0.45	0.26	0.19	0.18	
	urd	0.50	0.10	0.40	0.39	0.16	0.16	0.09	0.09	
Task	Lang	w_xlm	w_mmbert	w_mdeberta	t_ste	t_vil	t_deh	t_ext	t_lac	t_inv
3	amh	0.85	0.05	0.10	0.49	0.55	0.44	0.45	0.43	0.47
	ben	0.85	0.00	0.15	0.3	0.39	0.35	0.45	0.26	0.40
	deu	0.40	0.15	0.45	0.35	0.19	0.17	0.33	0.23	0.22
	fas	0.75	0.05	0.20	0.4	0.46	0.49	0.49	0.51	0.34
	hau	1.00	0.00	0.00	0.43	0.23	0.36	0.29	0.41	0.30
	hin	0.65	0.15	0.20	0.41	0.39	0.33	0.49	0.43	0.52
	khm	0.55	0.05	0.40	0.62	0.28	0.28	0.36	0.46	0.32
	nep	0.60	0.15	0.25	0.48	0.43	0.44	0.56	0.40	0.43
	ori	0.90	0.00	0.10	0.35	0.27	0.66	0.40	0.20	0.40
	pan	0.75	0.10	0.15	0.41	0.43	0.40	0.42	0.37	0.36
	swa	0.20	0.05	0.75	0.57	0.55	0.38	0.42	0.47	0.49
	tel	0.55	0.05	0.40	0.32	0.45	0.31	0.41	0.5	0.40
	tur	0.50	0.00	0.50	0.46	0.42	0.37	0.37	0.49	0.40
	urd	0.60	0.00	0.40	0.21	0.24	0.2	0.24	0.17	0.24

Table 8: Model ensemble weights ("w_") and class decision thresholds ("t_") for all low-resource languages across three subtasks. Label abbreviations: pol=Political, rac=Racial, rel=Religious, gen=Gender, oth=Other; ste=Stereotype, vil=Vilification, deh=Dehumanization, ext=Extreme Language, lac=Lack of Empathy, inv=Invalidation.

F Confusion matrix for English and Persian predictions on test data

Table 9, Table 10 and Table 11 show the confusion matrix for English prediction results across the three Subtasks. Figure 1 presents the heatmap visualisation. In Subtask 1, the model achieves balanced precision and recall, indicating robust binary classification without significant bias toward either class. In Subtask 2, the model performance varies substantially across labels. Political polarisation has high recall (0.857), though precision suffers from moderate false positives (213). However, Gender and Other categories struggle severely probably due to heavy class imbalance in training data. In Subtask 3, the model detects Vilification (F1: 0.639) and Extreme Language (F1: 0.629) with high recall, though Extreme Language has substantial false positives (231). Lack of Empathy is the hardest to detect, with lowest precision (0.310).

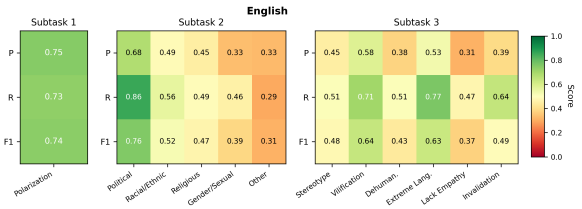


Figure 1: Heatmap of per-class Precision (P), Recall (R), and F1 for the English test predictions across all three subtasks. Each row corresponds to a class label; columns show the three metrics. Darker cells indicate higher values.

Table 9: English Subtask 1 confusion matrix and metrics.

Label	TN	FP	FN	TP	Precision	Recall
Polarization	788	131	142	391	0.749	0.734

Table 10: English Subtask 2 (Types): confusion matrix counts and positive-class metrics.

Label	TN	FP	FN	TP	Precision	Recall	F1
Political	721	213	74	444	0.676	0.857	0.756
Racial/Ethnic	1250	75	56	71	0.486	0.559	0.520
Religious	1370	31	26	25	0.446	0.490	0.467
Gender/Sexual	1389	30	18	15	0.333	0.455	0.385
Other	1359	35	41	17	0.327	0.293	0.309

Table 11: English Subtask 3 (Manifestations): confusion matrix counts and positive-class metrics.

Label	TN	FP	FN	TP	Precision	Recall	F1
Stereotype	1094	139	107	112	0.446	0.511	0.477
Vilification	888	189	110	265	0.584	0.707	0.639
Dehumanization	1128	148	86	90	0.378	0.511	0.435
Extreme Language	876	231	81	264	0.533	0.765	0.629
Lack of Empathy	1122	169	85	76	0.310	0.472	0.374
Invalidation	926	262	94	170	0.394	0.644	0.489

Table 12, Table 13 and Table 14 show the confusion matrix for Persian prediction results across the three Subtasks. Figure 2 presents the heatmap visualisation. In Subtask 1, the model achieves exceptional and balanced performance (precision = recall = 0.907) on the positive label. This reflects the high prevalence of polarised content in Persian data. In Subtask 2, Political, Religious and Gender/Sexual polarisation perform reasonably well with balanced precision-recall. However, Racial/Ethnic shows poor performance indicating the model struggles to learn this rare pattern despite reasonable precision-recall balance. This is probably due to lack of training data. In Subtask 3, Vilification detection is outstanding with near-perfect recall (0.939), consistent with its high prevalence in training data. Stereotype shows moderate recall but poor precision, suggesting over-prediction.

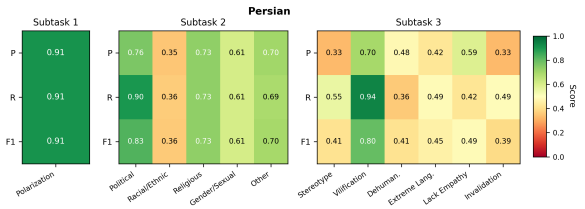


Figure 2: Heatmap of per-class Precision (P), Recall (R), and F1 for the Persian (Farsi) test predictions across all three subtasks. Each row corresponds to a class label; columns show the three metrics. Darker cells indicate higher values.

Table 12: Persian Subtask 1 confusion matrix counts and positive-class metrics.

Label	TN	FP	FN	TP	Precision	Recall
Polarization	283	102	102	997	0.907	0.907

Table 13: Persian Subtask 2 (Types): confusion matrix counts and positive-class metrics.

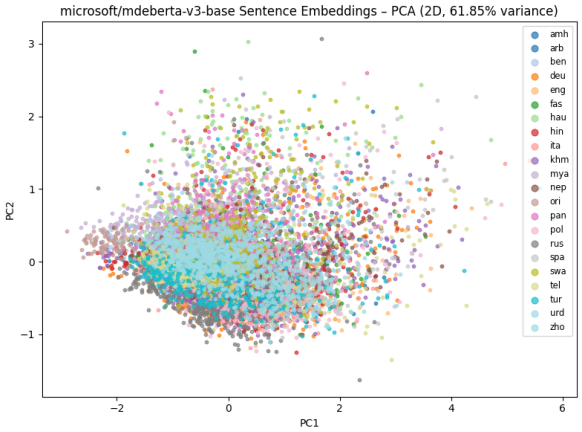
Label	TN	FP	FN	TP	Precision	Recall	F1
Political	650	183	64	587	0.762	0.902	0.826
Racial/Ethnic	1424	24	23	13	0.351	0.361	0.356
Religious	1303	38	38	105	0.734	0.734	0.734
Gender/Sexual	1360	35	35	54	0.607	0.607	0.607
Other	1021	104	112	247	0.704	0.688	0.696

Table 14: Persian Subtask 3 (Manifestations): confusion matrix counts and positive-class metrics.

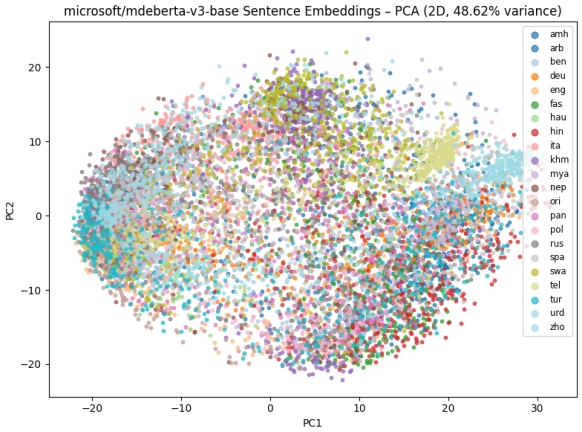
Label	TN	FP	FN	TP	Precision	Recall	F1
Stereotype	1068	221	87	108	0.328	0.554	0.412
Vilification	264	353	53	814	0.698	0.939	0.800
Dehumanization	1395	25	41	23	0.479	0.359	0.411
Extreme Language	1069	165	128	122	0.425	0.488	0.454
Lack of Empathy	1295	43	85	61	0.586	0.418	0.488
Invalidation	1247	119	60	58	0.328	0.492	0.393

G PCA analysis results

We randomly sample 500 sentences from the train data then encode texts with a model using the final-layer [CLS] embedding as the sentence representation. Results shown in Figure 3, Figure 4, Figure 5.

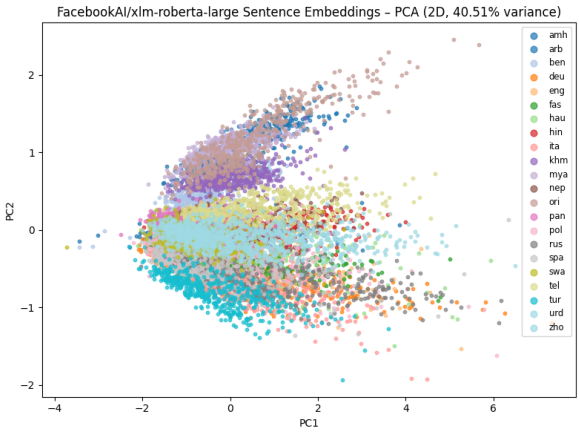


(a) mDeBERTa before fine-tuning.

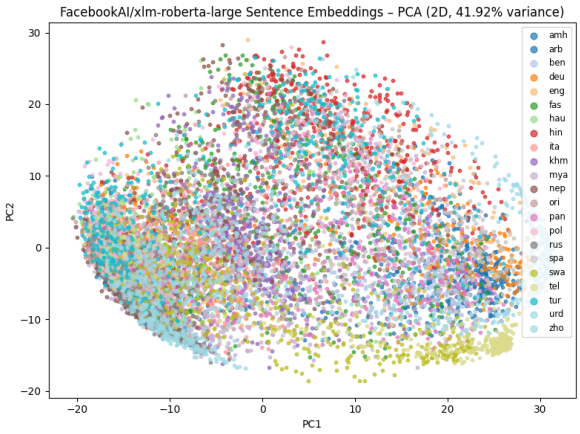


(b) mDeBERTa after fine-tuning (Subtask 2).

Figure 3: 2-D PCA projections of [CLS] embeddings for 500 randomly sampled training posts, encoded by mDeBERTa-v3-base. Each point represents one post, and colours denote the language of the post. (a) Pre-trained model before fine-tuning: posts from different languages overlap densely, suggesting a strongly language-agnostic shared space. (b) After fine-tuning on Subtask 2: the geometry remains globally mixed across languages while becoming more spread out, indicating that task-specific fine-tuning further aligns cross-lingual representations.

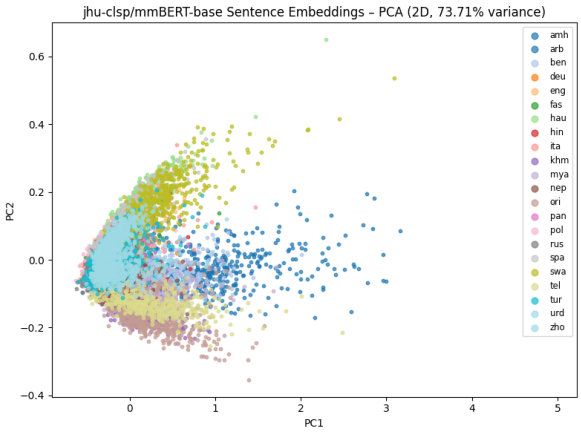


(a) XLM-R before fine-tuning.

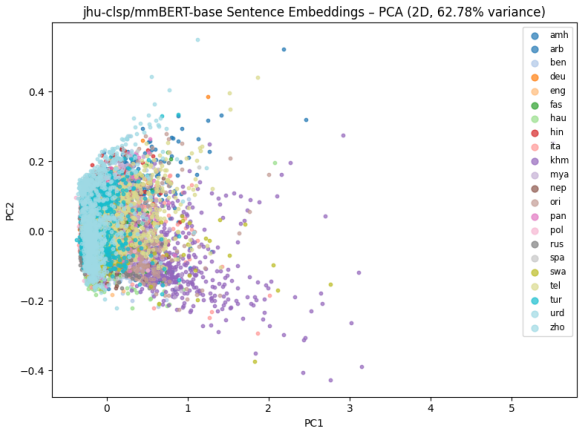


(b) XLM-R after fine-tuning (Subtask 2).

Figure 4: 2-D PCA projections of [CLS] embeddings for 500 randomly sampled training posts, encoded by XLM-RoBERTa-Large. Each point is one post; colours denote the language of the post. (a) Pre-trained model: languages overlap broadly, with several languages denser in particular regions. (b) After Subtask 2 fine-tuning: representations become more globally spread and intermixed, suggesting improved cross-lingual alignment for the task.



(a) mmBERT before fine-tuning.



(b) mmBERT after fine-tuning (Subtask 2).

Figure 5: 2-D PCA projections of [CLS] embeddings for 500 randomly sampled training posts, encoded by mmBERT-base. Each point is one post; colours denote the language of the post. (a) Pre-trained model: clear language-specific clusters and a long-tail structure dominated by a subset of languages, indicating that language identity remains salient in the embedding space. (b) After Subtask 2 fine-tuning: clusters partially merge but language-specific structure is still visible, suggesting that fine-tuning does not fully wash out language-identity signals in mmBERT.

H Computational Cost and Scalability

Our pipeline is computationally non-trivial due to the combination of cross-validation, multiple backbones, and multiple seeds across 22 languages and three subtasks. In total, we performed approximately 3600 fine-tuning runs: for each language–subtask pair, we trained 3 multilingual encoders (mDeBERTa, XLM-R, mmBERT) with 3 random seeds across 5 cross-validation folds, plus an additional monolingual encoder for the 6 high-resource languages, followed by a final full-data fine-tuning stage. Each run takes a few minutes on a single NVIDIA A100 GPU, depending on backbone size and language-specific dataset size. Inference is substantially cheaper: forward-passing the full test set for a single model takes well under a minute per language, and ensembling adds only logit averaging and thresholding, which are negligible.

For real-world deployment, the cross-validation and grid-search stages are a one-time development cost per language. At inference time, the deployed system requires only the seed-averaged ensemble (1–2 models per language) and the precomputed thresholds, which is comparable in cost to a single fine-tuned encoder. Because each language is tuned independently, the pipeline parallelises trivially across GPUs and scales linearly with the number of languages. The main scalability bottleneck is the cross-validation phase, which could be reduced (e.g., 3 folds or fewer seeds) at the cost of slightly higher variance in the estimated thresholds and ensemble weights.

I Threshold Sensitivity

A natural concern is whether the tuned class-wise thresholds are brittle under small perturbations. Our threshold selection procedure is designed to mitigate this in two ways. First, thresholds are tuned on out-of-fold (OOF) logits aggregated across 5 folds and averaged across 3 random seeds, so each threshold is selected against a relatively large and noise-reduced calibration set rather than a single validation split. Second, we use a coarse optimisation step of 0.01, which discourages overfitting to spurious decimal-level optima on the calibration data. The substantial gap between fixed-threshold ($t = 0.5$) performance and tuned performance, together with the wide cross-lingual variation in optimal thresholds (ranging from 0.02 to 0.99 in Subtask 1; see Appendix E), indicates that the macro-F1 landscape is genuinely shaped by class priors rather than by narrow, unstable peaks. We would therefore expect modest threshold perturbations (e.g., ± 0.05) to produce small rather than catastrophic changes in macro-F1 for most languages, with greater sensitivity for languages whose optimal thresholds lie near the extremes (e.g., Khmer with $t = 0.99$ in Subtask 1, or Hausa with $t = 0.02$), where any movement toward 0.5 effectively collapses the prediction to the majority class. A systematic perturbation study quantifying this sensitivity is left as future work.

J Model Selection and Ablation Discussion

Model selection methodology. We selected backbones through a two-stage probing procedure on the development set. For multilingual encoders, we evaluated three strong publicly available models that span complementary pretraining recipes: XLM-RoBERTa-Large (massively multilingual, contrastive), mDeBERTa-v3 (disentangled attention, strong monolingual transfer), and mmBERT (annealed language learning with broader low-resource coverage). All three were retained in the ensemble pool because no single backbone consistently dominated across languages and subtasks (see Appendix E). For monolingual encoders, we restricted attention to the 6 high-resource languages where well-pretrained, domain-appropriate checkpoints exist, and selected the best-performing checkpoint per language based on dev-set macro-F1 (Table 5). Languages without a clearly superior monolingual model were treated under the multilingual-only setting.

Implicit ablation from ensemble weights. Although we do not include a full controlled ablation, the tuned ensemble weights in Appendix E provide implicit ablation evidence. For high-resource languages, the optimisation consistently assigns the majority weight to the monolingual encoder ($w_{\text{mono}} \in [0.55, 0.90]$), confirming that monolingual specialisation contributes meaningfully on top of the multilingual signal. Conversely, the multilingual weight is never reduced to zero, indicating that multilingual encoders provide complementary information rather than being redundant. For low-resource languages, the weight assigned to each multilingual backbone varies substantially across languages and subtasks (e.g., XLM-R dominates for Bengali and Odia in Subtask 1, while mDeBERTa dominates for Swahili), showing that no

single multilingual backbone is uniformly optimal and that the ensemble adapts language by language. The fact that the optimisation rarely collapses to a single model (i.e., a one-hot weight vector) is itself evidence that ensembling contributes beyond any individual backbone.

Implicit ablation from threshold tuning. The wide cross-lingual variation in tuned thresholds (e.g., 0.02 for Hausa, 0.99 for Khmer in Subtask 1) provides further implicit ablation evidence: a uniform threshold of 0.5 would be far from optimal for many languages and labels, and would systematically collapse predictions for languages with extreme class priors. This indicates that the threshold-tuning component is not merely cosmetic but structurally necessary for macro-F1 optimisation in this multilingual, imbalanced setting. A more systematic controlled ablation (single-backbone vs. ensemble, fixed vs. tuned thresholds) is left as future work.

K Sample predictions for error analysis

Table 15 provides representative examples of correctly classified and missed test instances for English and Persian across the three subtasks, supplementing the qualitative error analysis in Section 4.2 (Error Analysis). Persian phrases are shown in the original script with English glosses. "–" in the Predicted column denotes a missed (false-negative) prediction for the corresponding label.

Example (excerpt)	Subtask	Gold	Predicted
English – Cases correctly detected			
Post containing slurs and explicit outgroup labelling (e.g., “nazis”, “fascist”)	S1: Detection	Polarised	Polarised
Post directly naming a political ideology (e.g., “far right”)	S2: Type	Political	Political
Post directly naming a religion (e.g., “Christian”)	S2: Type	Religious	Religious
Post calling for extreme action (“deport to x”, “ruin x”)	S3: Manifestation	Lack of Empathy	Lack of Empathy
English – Cases missed			
Sarcastic rhetorical question with implicit immigration framing	S1: Detection	Polarised	Not Polarised
Indirect ideological reference (“blue/red states”)	S2: Type (Religious)	Religious	–
“Reasonable-sounding” commentary correcting someone’s “misbelief”	S3: Manifestation	Lack of Empathy	–
Persian – Cases correctly detected			
اخوند وطن فروش (“cleric”, “traitor”) – explicit hostile labelling	S1: Detection	Polarised	Polarised
اعدام باید گردد (“must be executed”) – explicit endorsement of punishment	S1: Detection	Polarised	Polarised
Post with overt national / ethnic markers	S2: Type	Race	Race
Post explicitly referencing women’s identity	S2: Type	Gender	Gender
اعدام آتش بزن (“execution”, “set on fire”) – explicit call for violence	S3: Manifestation	Lack of Empathy	Lack of Empathy
Persian – Cases missed			
References to زندان اوین اعتصاب ها فیلترینگ (Evin prison, strikes, filtering) – implicit framing via topic selection	S1: Detection	Polarised	Not Polarised
Critique of nationality / migration / cultural norms without explicit naming	S2: Type (Race)	Race	–
Descriptive commentary or sarcasm that trivialises harm without explicit violence cues	S3: Manifestation	Lack of Empathy	–

Table 15: Representative examples of correctly classified and missed test instances for English and Persian across the three subtasks. Persian phrases are shown in the original script with English glosses; English examples include short excerpts and characteristic phrases.