NASIM_Lab at SemEval-2026 Task 9: A Comparative Analysis of Fine-Tuned Small Language Models vs. Generative Large Language Models for Multilingual Polarization Type Detection

Neel Sabhahit*, **Sanjeevan Selvaganapathy*** **and Mehwish Nasim**†
Network Analysis and Social Influence Modelling (NASIM) Lab
School of Physics, Mathematics and Computing
The University of Western Australia
{neel.sabhahit,sanjeevan.selvaganapathy,mehwish.nasim}@uwa.edu.au

Abstract

The POLAR dataset contains various social media texts that might be polarized (conflict-inducing or dangerously divisive). The task at hand is to identify whether any of the following types of polarization are present: political, racial/ethnic, religious, gender/sexual, and other types across 22 languages. In this paper, we propose a system of fine-tuned language-specific small language models and compare our approach with state-of-the-art large language models on the POLAR dataset. By fine-tuning models for each language, we demonstrate that fine-tuned small encoder-only models consistently outperform large language models, especially for low-resource languages. Our system [1] performs well on this task for most low-resource languages, notably taking the top spot on the leaderboard in Burmese (mya), appearing within the top 10 for 12 languages, and within the top 20 for all remaining languages.

1 Introduction

Polarization is defined as the process of formation of population clusters or groups where attributes within clusters are similar but differ significantly from attributes in other clusters. These differences often lead to tensions and conflict between groups in society (Esteban and Ray, 1994). As noted by Kish Bar-On et al. (2024), polarization lies across multiple dimensions, such as political, social, and cultural differences, and causes deeper political divides, disrupts society, and weakens institutional trust. Some studies (Van Bavel et al., 2021; Kubin and Von Sikorski, 2021) have noted that while social networks, such as Twitter and Facebook, can play a role in amplifying polarization by creating echo chambers and divisive content, there are also conflicting observations that claim the opposite. Regardless of these conflicting findings, systematically identifying polarizing topics remains a critical prerequisite for understanding and mitigating polarization dynamics.

The POLAR dataset, which has been used as the source for our analysis, is a multilingual corpus consisting of 22 languages (Naseem et al., 2026b). The shared task, SemEval-2026 Task 9, consists of 3 subtasks: (1) to identify the presence of polarization; (2) to identify the types of polarization; and (3) to identify the target groups of polarization (Naseem et al., 2026a). We focused on Subtask 2, which involves the classification of social media text across five polarization dimensions: political, racial/ethnic, religious, gender/sexual, and other.

The system we present utilizes language-specific model selection by fine-tuning small encoder-only transformer models for each language separately. Our approach ranked within the top 10 on the leaderboard for 12 languages (including first in Burmese, third in Swahili, fifth in Hindi, and sixth in Punjabi), and also performed well on the remaining languages, placing within the top 20 submissions on the leaderboard. In addressing this task, we faced two primary challenges: (1) *LLM content filtering mechanisms* and (2) *the availability of limited contextual information* in many of the texts.

Through our analysis, we demonstrate that fine-tuned small encoder-based models consistently outperform state-of-the-art large language models in low-resource settings. Our results highlight the effectiveness of smaller, specialized architectures for multilingual polarization detection.

2 Background

Large language models (hereafter referred to as LLMs) and fine-tuned small language models (hereafter referred to as SLMs), both built on the transformer architecture (Vaswani et al., 2017), have

*Equal contribution
†Project advisor
[1]Code available at github.com/sanjerine/polar-semeval-nasim-lab

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3316–3327

been used extensively in the detection and analysis of polarization and hate speech (Chowdhury et al., 2026; Wu and Hsun, 2025).

In our study, we build on the work of Naseem et al. (2026b), who evaluated the effectiveness of various fine-tuned SLMs and LLMs on the POLAR dataset and reported lower performance of both model families on the second and third subtasks of the POLAR benchmark.

We also compare the effectiveness of LLMs and fine-tuned SLMs on both low- and high-resource languages. As defined by Magueresse et al. (2020), low-resource languages (henceforth referred to as LRLs) are languages for which data is scarce and no statistical methods can be reliably applied. We use the classifications provided in the No Language Left Behind initiative (NLLB Team et al., 2022) to separate the languages in the dataset into low- and high-resource groups for analysis. Amharic, Burmese, Hausa, Khmer, Nepali, Odia, Punjabi, Telugu, and Urdu are all defined as low-resource, while the rest of the languages in our dataset are defined as high-resource.

3 System Overview

3.1 Model Selection and Evaluation

To determine the optimal model configuration, we evaluated various LLMs (both proprietary and open-source) and open-source fine-tuned SLMs. During model evaluation, we encountered two primary issues: content filtering constraints in certain proprietary LLMs and the inability to use some smaller models due to restricted access to model weights and incomplete documentation. The evaluation metric used for the task was the Macro-F1 metric, which is defined as the average of F1 scores per label across all labels. More formally,

$$\text{Macro-F1} = \frac{1}{L}\sum_{l=1}^{L} \text{F1}^{(l)} \tag{1}$$

where L is the number of labels in the dataset considered (in our case, $L = 5$ for Subtask 2).

As the task was formulated as a multi-label classification problem, Binary Cross Entropy (BCE) with logits was used as the loss function to fine-tune most models, defined for a single training instance as:

$$\mathcal{L}_{BCE} = -\frac{1}{L}\sum_{i=1}^{L} \begin{cases} \log(\hat{y}_i) & y_i = 1 \\ \log(1-\hat{y}_i) & y_i = 0 \end{cases} \tag{2}$$

For the XLM-R models, we used Asymmetric Loss (Ridnik et al., 2021) as it provided improved performance on the development set compared to BCE loss. The Asymmetric Loss for a single training instance can be given as:

$$\mathcal{L}_{ASL} = -\frac{1}{L}\sum_{i=1}^{L} \begin{cases} (1-p_i)^{\gamma_+}\log(p_i) & y_i = 1 \\ (p_i)^{\gamma_-}\log(1-p_i) & y_i = 0 \end{cases} \tag{3}$$

The specific hyperparameter values for γ_+ and γ_- can be found in Appendix A.2.

3.2 Fine-tuned SLMs[2]

We developed a system utilizing a language-specific model selection strategy with encoder-only SLMs fine-tuned for each language. This approach allowed us to leverage specialized architectures optimized for each language, moving away from a general multilingual encoder for all languages. All of the chosen models were further fine-tuned on the POLAR dataset for Subtask 2.

For the majority of languages in our dataset, we utilized the XLM-RoBERTa architecture (Conneau et al., 2020). This included Bengali, German, English, Persian, Hindi, Italian, Khmer, Burmese, Nepali, Odia, Punjabi, Telugu, Turkish, and Urdu.

For the remaining languages, we used specialized regional models. We used models with architectures based on BERT (Devlin et al., 2019), such as ruRoBERTa (Zmitrovich et al., 2024) for Russian and MARBERTv2 (Abdul-Mageed et al., 2021) for Arabic. Similarly, for Polish and Spanish, we used models based on RoBERTa (Liu et al., 2019) that had been continually pretrained on Polish (Dadas et al., 2020) and Spanish (De la Rosa et al., 2022) corpora, respectively.

For languages in the African subcontinent such as Amharic, Hausa and Swahili, we utilized AfroXLMR (Alabi et al., 2022). Although our final submission also included a model for Chinese, we have omitted its description in this paper due to the identification of several model-related errors post-submission. The huggingface identifiers of the models used are provided in Appendix A.1.

3.3 Large Language Models

Various LLMs were also evaluated across proprietary and open-source families. These models were not officially submitted, as they frequently underperformed relative to our fine-tuned SLMs on the

[2]Our official submission

test set. For the purpose of benchmarking, we selected the LLMs that demonstrated the highest average Macro-F1 score across all languages on the development set, which is defined as

$$\text{Macro-F1}_{Average} = \frac{1}{L}\sum_{\ell=1}^{L}\text{Macro-F1}_{\ell} \quad (4)$$

where L is the number of languages evaluated (in our case, $L = 21$).

4 Experimental Setup

4.1 Fine-tuned SLMs

The fine-tuning process was implemented using the Huggingface Trainer API (Wolf et al., 2020). We used NVIDIA A100 GPUs to accelerate the fine-tuning process and accommodate the memory requirements of larger transformer architectures.

For the first phase of the official task (the development phase), we utilized only the official training dataset. To ensure robust model selection and generalization, we performed a 90/10 internal training-validation split. In this phase, the official development set was used as an internal test set to evaluate and select the best model configuration per language.

In the second phase (evaluation phase), the official test set was used to evaluate the quality of our submissions. For the final submission, a combination of the official training and development sets was used to fine-tune the models. This enabled maximum utilization of the available training data, as the models were exposed to a wider range of label combinations and contexts.

Lang	Rank	Lang	Rank	Lang	Rank
mya	1	pol	7	arb	13
swa	3	khm	8	spa	13
hin	5	nep	7	ita	15
pan	6	hau	8	tel	17
amh	7	ben	13	deu	20
rus	7	fas	13	eng	19
tur	7	ori	7	urd	20

Table 1: Official leaderboard rankings of our system across all languages, excluding Chinese.

The hyperparameters chosen for each model can be found in Appendix A.2. To convert the models' sigmoid output probabilities into binary labels, we tuned decision thresholds separately for each language–label pair via grid search on the training data. We evaluated thresholds $t \in \{0.05, 0.10, \ldots, 0.90\}$ (step size 0.05) using 5-fold cross-validation and selected the value that maximized the mean F1 score for that label. In cases where a specific label lacked positive training instances for a given language, we used the default threshold $t = 0.5$.

4.2 Large Language Models

Inference for large language models was performed using the OpenRouter API, which routes requests to multiple LLM providers. Three prompt families, Zero-shot, Few-shot (Brown et al., 2020) and Zero-shot Chain-of-thought (Wei et al., 2022; Kojima et al., 2022) were evaluated as user prompts across five models: Gemini-3-Flash-Preview (hereafter referred to as Gemini 3 Flash) (Google, 2026), Grok 4.1-Fast (hereafter referred to as Grok 4.1 Fast) (xAI), GPT 4o-mini (OpenAI), Qwen3.5-397B-A17B (hereafter referred to as Qwen 3.5 397B) (Qwen Team, 2026) and DeepseekV3.2 (DeepSeek-AI, 2025).

The above models are all current state-of-the-art language models, and were selected to maximize diversity in architecture and scale. The top three model-prompt combinations were selected based on evaluation on the development set using average Macro-F1 scores.

The selected models were then evaluated on the test set. Further details on LLM inference are provided in Appendix A.3, and all prompts used are provided in Appendix A.4.

5 Results

5.1 Fine-tuned SLMs

The fine-tuned models demonstrate competitive performance across the official leaderboard for many languages, as seen in Table 1. Notably, our system achieved the top rank for Burmese (mya), third in Swahili (swa), fifth in Hindi (hin) and sixth in Punjabi (pan). We were in the top 10 ranks for Amharic (amh), Russian (rus), Turkish (tur), Odia (ori), Polish (pol), Khmer (khm), Nepali (nep) and Hausa (hau). For the remaining nine languages, we placed between 13th and 20th. Table 2 presents the Macro-F1 scores obtained by the fine-tuned SLMs on the official test set, alongside other models that we evaluated.

The average Macro-F1 score across high-resource languages was 0.5336 and across low-resource languages was 0.5833, indicating that our

Language	Fine-tuned SLMs	Gemini 3 Flash	Grok 4.1 Fast	Qwen 3.5 397B
amh	**0.6059**	0.5044	0.4337	0.4749
arb	0.6047	**0.6136**	0.5913	0.5558
ben	0.3240	**0.3371**	0.3141	0.3225
mya	**0.7474**	0.6212	0.4645	0.4808
eng	0.4581	**0.5099**	0.4647	0.4613
deu	0.5049	0.5346	**0.5409**	0.4934
hau	**0.3764**	0.1861	0.1567	0.1800
hin	**0.7846**	0.5368	0.5930	0.4911
ita	0.2463	**0.5806**	0.5440	0.5171
khm	**0.6745**	0.1284	0.2175	0.1422
nep	**0.7754**	0.7653	0.7097	0.6964
ori	**0.5549**	0.4288	0.4873	0.4468
fas	**0.5793**	0.3722	0.4384	0.3778
pol	0.5710	**0.5938**	0.5550	0.5340
pan	0.4785	**0.5659**	0.4397	0.4242
rus	**0.5745**	0.5357	0.4533	0.5208
spa	**0.6231**	0.5950	0.5270	0.5126
swa	**0.5360**	0.3808	0.3388	0.3338
tel	**0.3049**	0.1554	0.2258	0.2024
tur	0.5962	**0.6583**	0.6090	0.5832
urd	**0.7314**	0.3756	0.4110	0.3334

Table 2: Comparison of Macro-F1 scores on the official test set across different languages and model architectures, with LLMs evaluated using few-shot prompts.

language-specific fine-tuning strategy was effective in low-resource settings.

5.2 Large Language Models

The final test set evaluation of LLMs was conducted using Gemini 3 Flash, Grok 4.1 Fast, and Qwen 3.5 397B, as they achieved the highest average Macro-F1 scores on the development set. We also exclusively used few-shot prompts in this stage, as it consistently performed better than the other prompt families.

Model	Langs. with Top Score
Gemini 3 Flash	**14**
Grok 4.1 Fast	7
Total	**21**

Table 3: Distribution of top-performing LLMs across the 21 target languages on the official test set.

As seen in Table 3, Gemini 3 Flash emerged as the best model, achieving the highest Macro-F1 scores across 14 languages. Grok 4.1 Fast achieved the best Macro-F1 score across the remaining seven languages.

The average Macro-F1 score varied significantly between language groups, with Gemini 3 Flash achieving the best score of 0.5207 in high-resource languages, followed by Grok 4.1 Fast with 0.4975 and Qwen 3.5 397B with 0.4753. For low-resource languages, Gemini 3 Flash had an average Macro-F1 of 0.4146, followed by Grok 4.1 Fast with 0.3940 and Qwen 3.5 397B with 0.3757. The significant gap between the average Macro-F1 scores of the low- and high-resource groups indicates that the performance of LLMs degrades in low-resource settings. Furthermore, this gap is more pronounced in LLMs than in our fine-tuned SLMs.

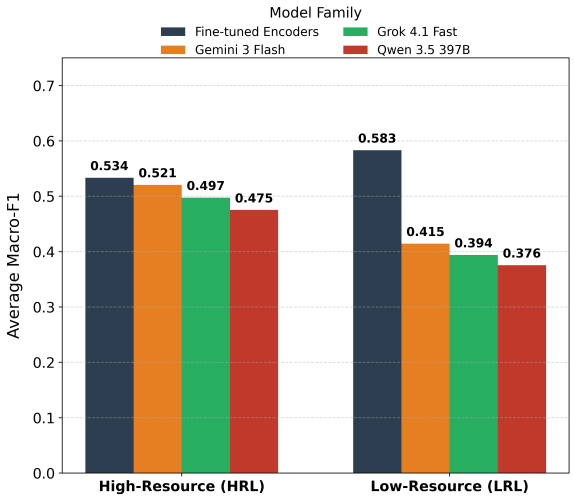

Figure 1: Comparison of Average Macro-F1 scores for the fine-tuned models and LLMs across high- and low-resource groups on the official test set.

5.3 Comparison between Fine-tuned SLMs and LLMs

Comparative analysis shows that our system consistently outperforms the LLMs across nearly all

LRLs, with the exception of Punjabi, where Gemini 3 Flash achieves a better Macro-F1 score. Figure 1 shows the average Macro-F1 scores of our system and the LLMs on the official test set for both resource groups. While our system achieves higher Macro-F1 scores on Hindi, Persian, Russian, Spanish, and Swahili within the high-resource group, Grok 4.1 Fast outperforms it on German, and Gemini 3 Flash on the remaining high-resource languages.

Label Category	Average F1 Score
Political	0.6835
Religious	0.6166
Racial/Ethnic	0.5294
Gender/Sexual	0.5225
Other	0.4223

Table 4: Average Macro-F1 Scores across all 21 Languages per polarization category for the fine-tuned SLMs on the official test set.

5.4 Observations

Difficulty in predicting the "other" category: From Table 4, we observe that the average Macro-F1 score across all languages for the "other" label is 0.4223. This is significantly lower than the other categories and suggests that our system does not do well at classifying polarization types that lack explicit definitions. We also notice that while the F1 score for the "other" category is high for languages such as Khmer (0.8895) and Burmese (0.8421), it falls significantly for languages such as Hausa (0.0000), Italian (0.0439), and English (0.2093). This volatility indicates an inconsistent pattern: while our system captures implicit polarization patterns in some cultural contexts/languages, it fails to capture these nuances in other languages, despite some languages such as English and Burmese using models with the same underlying architecture. Further study and analysis[3] may help understand these nuances better.

6 Conclusion

In this study, we have developed a system consisting of 21 specialized models, fine-tuned for each language on the POLAR dataset for Subtask 2. By fine-tuning language-specific encoder models, we have ensured that our system obtains the optimal Macro-F1 score per language. We have also established that fine-tuned small language models consistently outperform general large language models in classifying polarization in low-resource languages.

Future work may extend this work by evaluating the efficacy of fine-tuned language-specific large language models on the POLAR dataset, to understand the effects of scale on polarization. It would also be interesting to observe if the use of reasoning models can overcome the shortcomings of the LLMs used here.

By demonstrating the dominant performance of fine-tuned SLMs over LLMs in classifying polarization types across low-resource languages, we suggest a resource-efficient pathway for research communities lacking the capacity to train large-scale models to understand and improve polarization and hate-speech classification.

7 Ethical Considerations

We acknowledge that the content of this dataset and task may be distressing, and that we have ensured that anyone who has come into contact with the dataset during the course of the study received adequate support.

Our work is intended solely for research purposes in the shared-task setting and should not be deployed without appropriate human oversight and safeguards.

[3]Further analysis is deferred to future work due to the constraints of the shared task timeline

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jawad Chowdhury, Rezaur Rashid, and Gabriel Terejanu. 2026. Measuring Social Media Polarization Using Large Language Models and Heuristic Rules. In Aijun An, Alfredo Cuzzocrea, and Hongxin Hu, editors, *Social Networks Analysis and Mining*, volume 16324, pages 429–444. Springer Nature Switzerland, Cham.

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.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training Polish Transformer-Based Language Models at Scale. In *Artificial Intelligence and Soft Computing*, pages 301–314, Cham. Springer International Publishing.

Javier De la Rosa, Eduardo Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural*, pages 13–23.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Joan-Maria Esteban and Debraj Ray. 1994. On the Measurement of Polarization. *Econometrica*, 62(4):819.

Google. 2026. Gemini 3 Developer Guide. https://ai.google.dev/gemini-api/docs/gemini-3.

Kati Kish Bar-On, Eugen Dimant, Yphtach Lelkes, and David G Rand. 2024. Unraveling polarization: Insights into individual and collective dynamics. *PNAS Nexus*, 3(10):pgae426.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Emily Kubin and Christian Von Sikorski. 2021. The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, 45(3):188–206.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv preprint*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, 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)*, San Diego, CA, USA. 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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint*.

OpenAI. GPT-4o mini Technical Documentation. https://developers.openai.com/api/docs/models/gpt-4o-mini.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric Loss For Multi-Label Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, Montreal, QC, Canada. IEEE.

Jay J. Van Bavel, Steve Rathje, Elizabeth Harris, Claire Robertson, and Anni Sternisko. 2021. How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11):913–916.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shih-Hung Wu and Tsai Tsung Hsun. 2025. Fine-Tuned Models for Hate Speech Detection: Assessing Generalization on Social Media. In *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*, pages 250–255, San Jose, CA, USA. IEEE.

xAI. Grok 4.1 Fast Developer Documentation. https://docs.x.ai/developers/models/grok-4-1-fast-reasoning.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A Family of Pretrained Transformer Language Models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

A Appendix

A.1 SLM Fine-tuning

Table A.1 lists the Huggingface model IDs and their respective URLs. These are the models used in our official submission.

Table A.2 reports per-label F1 scores as well as the Macro-F1 scores for each language (as evaluated during the development phase of the task).

A.2 SLM Hyperparameters

All fine-tuning experiments were conducted using the following common configuration:

- **Optimizer:** AdamW with a weight decay coefficient of 0.01.
- **Learning Rate Schedule:** Linear decay with a warmup ratio of 0.1.
- **Sequence Length:** A maximum sequence length of 256 tokens.
- **Numerical Precision:** 16-bit floating-point (fp16) mixed precision.
- **Data Partitioning:**
  - 90/10 train–validation split.
  - Fixed random seed of 42 for reproducibility.
- **Model Selection:** Best checkpoints were selected based on the highest validation macro-F1 score.

A detailed list of models, configurations of hyperparameters and loss functions is available in Table A.3.

A.3 LLM Inference

Two instances in Swahili from the development set, with IDs swa_1497cbe557ff17b6046e198f41ea0379 and swa_43011832e0c4e351adc359e2fa6eda10, were not processed by Gemini 3 Flash due to API errors. We considered two evaluation scenarios: one assuming both predictions are correct, and one assuming both predictions are incorrect. We then calculated both optimistic and pessimistic Macro-F1 scores based on these cases.

All models were run with temperature set to 0 to ensure deterministic outputs. Furthermore, explicit reasoning was disabled where supported to avoid generating intermediate reasoning outputs. We also enforced a request timeout of 25 seconds to prevent slow requests from hanging execution.

Table A.4 reports the average Macro-F1 scores for each LLM-prompt combination across all languages on the development set. The best-performing models and prompt types, selected based on these scores, were then evaluated on the test set. Table A.5 presents the Macro-F1 scores of the top models from Table A.4: Gemini 3 Flash, Qwen 3.5 397B and Grok 4.1 Fast.

Both Table A.4 and A.5 use the optimistic Macro-F1 scores for Gemini 3 Flash on Swahili, where missing predictions are assumed to be correct.

A.4 LLM Prompts

Zero-shot (Figure A.1), few-shot (Figure A.2) and zero-shot chain-of-thought (Figure A.3) prompt families were used in this study. The zero-shot prompt serves as a baseline reflecting default model behaviour, while the few-shot and zero-shot chain-of-thought prompts guide model outputs by incorporating task examples, and encouraging stepwise reasoning respectively.

In this study, we do not study prompt variations or methods such as supervised fine-tuning of LLMs due to time constraints.

Language	Hugging Face Model ID	Model URL
amh	Davlan/afro-xlmr-large	https://huggingface.co/Davlan/afro-xlmr-large
arb	UBC-NLP/MARBERTv2	https://huggingface.co/UBC-NLP/MARBERTv2
ben	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
deu	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
eng	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
fas	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
hau	Davlan/afro-xlmr-large	https://huggingface.co/Davlan/afro-xlmr-large
hin	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
ita	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
khm	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
mya	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
nep	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
ori	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
pan	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
pol	sdadas/polish-roberta-large-v2	https://huggingface.co/sdadas/polish-roberta-large-v2
rus	ai-forever/ruRoberta-large	https://huggingface.co/ai-forever/ruRoberta-large
spa	bertin-project/bertin-roberta-base-spanish	https://huggingface.co/bertin-project/bertin-roberta-base-spanish
swa	Davlan/afro-xlmr-large	https://huggingface.co/Davlan/afro-xlmr-large
tel	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
tur	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large
urd	FacebookAI/xlm-roberta-large	https://huggingface.co/FacebookAI/xlm-roberta-large

Table A.1: Model identifiers and Huggingface repository links for the 21 languages evaluated in this study.

Lang.	Dev Macro-F1	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
amh	0.5129	0.8711	0.5992	0.6296	0.3333	0.5964
arb	0.5800	0.7306	0.6188	0.6316	0.5939	0.4484
ben	0.4349	0.7251	0.0000	0.2449	0.3333	0.3166
deu	0.5295	0.6151	0.5172	0.5714	0.6099	0.2109
eng	0.3402	0.7133	0.4306	0.5773	0.3600	0.2093
fas	0.6171	0.8235	0.1481	0.6667	0.5799	0.6782
hau	0.5039	0.5564	0.5800	0.3750	0.3704	0.0000
hin	0.8075	0.9191	0.8112	0.9360	0.7722	0.4844
ita	0.4257	0.0234	0.4370	0.5846	0.1429	0.0439
khm	0.6945	0.7714	0.3684	0.7150	0.6279	0.8895
mya	0.5703	0.8426	0.7465	0.6154	0.6905	0.8421
nep	0.7818	0.7129	0.8716	0.8828	0.8085	0.6010
ori	0.5822	0.6918	0.5510	0.6777	0.4364	0.4179
pan	0.4501	0.6598	0.3368	0.5152	0.6122	0.2687
pol	0.6280	0.7694	0.5455	0.6479	0.5376	0.3548
rus	0.6378	0.6993	0.6189	0.7107	0.6127	0.2308
spa	0.5896	0.6719	0.5378	0.5987	0.7688	0.5383
swa	0.4637	0.4868	0.7795	0.8219	0.2752	0.3167
tel	0.4253	0.4695	0.2723	0.1132	0.2619	0.4076
tur	0.6391	0.7390	0.6343	0.7088	0.5608	0.3382
urd	0.7707	0.8621	0.7124	0.7245	0.6843	0.6736

Table A.2: SLM performance on the development set and per-label F1 scores for each language, as evaluated during the development phase of the task.

Model & Loss Function	Training Configuration
XLM-R	
Asym. ($\gamma_- = 4, \gamma_+ = 1$, clip=0.05)	Ep: 5, BS: 32, GA: 1, LR: 2e-5
AfroXLMR large	
BCE	Ep: 10, BS: 16, GA: 1, LR: 3e-5
MARBERTv2	
BCE	Ep: 10, BS: 16, GA: 1, LR: 2e-5
Polish RoBERTa large	
BCE	Ep: 10, BS: 16, GA: 1, LR: 2e-5
ruRoBERTa	
BCE	Ep: 10, BS: 16, GA: 1, LR: 2e-5
BERTIN RoBERTa base Spanish	
BCE	Ep: 10, BS: 16, GA: 1, LR: 2e-5

Table A.3: Model-specific hyperparameters. Ep: training epochs; BS: batch size; GA: gradient accumulation steps; LR: learning rate.

Model	Zero-shot	Few-shot	Chain-of-Thought
Gemini 3 Flash	0.465	**0.480**	0.462
Grok 4.1 Fast	0.427	**0.447**	0.428
Qwen 3.5 397B	0.433	**0.441**	0.435
GPT-4o-mini	0.409	0.401	0.416
DeepSeek V3.2	0.408	0.379	0.433

Table A.4: Average Macro-F1 across all languages per LLM-prompt combination on the development set.

Language	gemini 3 flash	qwen 3.5 397b	grok 4.1 fast
amh	**0.4293**	0.3972	0.3496
arb	**0.5808**	0.5203	0.5689
ben	**0.3259**	0.3030	0.2871
deu	**0.5444**	0.5079	0.5184
eng	**0.4913**	0.4578	0.4731
fas	0.4451	0.4606	**0.4768**
hau	0.1982	**0.2254**	0.1879
hin	0.6780	0.5792	**0.6942**
ita	0.3434	**0.3440**	0.3430
khm	0.5957	**0.6856**	0.5336
mya	**0.5715**	0.4197	0.3391
nep	**0.7783**	0.7661	0.7012
ori	0.5390	**0.5426**	0.4616
pan	**0.6211**	0.5149	0.4679
pol	**0.7237**	0.6259	0.5836
rus	0.5275	**0.5618**	0.5407
spa	0.5820	**0.5855**	0.5739
swa	**0.4131**	0.3369	0.4056
tel	**0.2970**	0.2497	0.2548
tur	**0.6891**	0.5292	0.6062
urd	0.3996	0.3563	**0.4054**

Table A.5: Comparative performance (best Macro-F1 scores) of Gemini 3 Flash, Grok 4.1 Fast and Qwen 3.5 397B over the development set.

```
ZERO_SHOT_PROMPT = """Classify this text for polarization types.

Text: {text}

For each label, output 1 if present, 0 if not.
Labels: political, racial/ethnic, religious, gender/sexual, other

JSON only, no explanation:
{{"political": 0, "racial/ethnic": 0, "religious": 0, "gender/sexual": 0, "other": 0}}"""
```

Figure A.1: The zero-shot prompt used for all LLMs.

```
FEW_SHOT_PROMPT = """Classify text for polarization. Critical distinction:
- Polarized: Contains divisive rhetoric, vilification, stereotyping, or us-vs-them framing
- NOT polarized: Neutral reporting, factual statements, or balanced discussion - even on
    controversial topics

<examples>
Text: "They claim the students were killed by Ukrainian Nazis"
Analysis: Reporting what others claim. Neutral framing.
{{"political": 0, "racial/ethnic": 0, "religious": 0, "gender/sexual": 0, "other": 0}}

Text: "Keep up the good fight zelensky ukraine"
Analysis: Expression of support. No divisive rhetoric.
{{"political": 0, "racial/ethnic": 0, "religious": 0, "gender/sexual": 0, "other": 0}}

Text: "Its the eternal narrative, because it dovetails so neatly with the fear, racism, and
    xenophobia theyre peddling."
Analysis: Accuses others of peddling fear and racism. Divisive framing.
{{"political": 1, "racial/ethnic": 1, "religious": 0, "gender/sexual": 0, "other": 0}}

Text: "Poland helping Ukraine Whites helping Whites. Where was Poland for the Middle East Crisis?"
Analysis: Racial framing ("Whites helping Whites"), implies racial bias in aid. Divisive.
{{"political": 1, "racial/ethnic": 1, "religious": 0, "gender/sexual": 0, "other": 0}}

Text: "OU community shares opinions on IsraelHamas war"
Analysis: Neutral news headline. No stance or divisive rhetoric.
{{"political": 0, "racial/ethnic": 0, "religious": 0, "gender/sexual": 0, "other": 0}}

Text: "Never call the fire fighters if your house burns down. Never use the roads outside your home.
    Never call the police for yourself. All socialism."
Analysis: Sarcastic political attack, vilifying those who critique socialism. Divisive.
{{"political": 1, "racial/ethnic": 0, "religious": 0, "gender/sexual": 0, "other": 0}}
</examples>

Text: {text}

JSON only, no explanation:
{{"political": 0, "racial/ethnic": 0, "religious": 0, "gender/sexual": 0, "other": 0}}"""
```

Figure A.2: The few-shot prompt used for all LLMs.

```
ZERO_SHOT_COT_PROMPT = """Analyze this text for polarization by examining its TACTICS, not just its
    topic.

Text: {text}

Think through these manifestations:

1. STEREOTYPE: Does it generalize characteristics to all members of a group, ignoring individual
    differences?
2. VILIFICATION: Does it defame or demonize a group through exaggeration, misrepresentation, or
    biased framing?
3. DEHUMANIZATION: Does it strip a group of human qualities (comparing to animals, objects, denying
    dignity)?
4. EXTREME LANGUAGE: Does it use absolutist terms ("always", "never", "worst") or us-vs-them framing
    ("we vs they")?
5. LACK OF EMPATHY: Does it marginalize or dismiss other perspectives without understanding?
6. INVALIDATION: Does it deny or reject the identity/existence of people or groups?

If ANY manifestation is present, identify which polarization TYPE it targets:
- political (parties, ideologies, government)
- racial/ethnic (race, ethnicity, nationality)
- religious (faith, religious groups)
- gender/sexual (gender identity, sexuality)
- other (class, generation, other groups)

Final output must be ONLY a JSON object containing a "reasoning" key for your brief analysis and the
    polarization keys:
{{"reasoning": "your brief analysis here", "political": 0, "racial/ethnic": 0, "religious": 0, "
    gender/sexual": 0, "other": 0}}"""
```

Figure A.3: The zero-shot chain-of-thought prompt used for all LLMs.