Aatman at SemEval-2026 Task 9: Transfer Learning for Multilingual Polarization Detection

Aatman Vaidya
University of Tübingen
aatman-vrundavan.vaidya@student.uni-tuebingen.de

Abstract

This paper describes our system for Subtask 1 of SemEval-2026 Task 9: POLAR, which focuses on multilingual polarization detection. The task is formulated as a binary classification problem across 22 languages drawn from diverse online platforms and real-world events. We investigate three complementary approaches: supervised fine-tuning of multilingual encoder-only transformer models, zero- and few-shot classification using large language models (LLMs), and transfer learning from related harmful language tasks such as hate speech, toxicity, abusive language, and gender-based violence.

Among the supervised models, mDeBERTa achieved the strongest baseline performance. Prompt-based methods with open-weight LLMs showed limited effectiveness, particularly in zero-shot settings. The best results were obtained using transfer learning, where the model was first fine-tuned on related task datasets and then adapted to the polarization task, achieving a Macro-F1 score of 0.781. Our findings indicate that supervised multilingual encoders remain highly effective for polarization detection and that incorporating related harmful language tasks can substantially improve performance, especially for nuanced and context-dependent expressions of polarization.

1 Introduction

Online polarization leads to hostility between social, political, or identity groups and has also been linked to real world violence (Zelalem and Guest, 2021). Polarized online discourse frequently contains harmful content, including hate speech, toxicity, sarcasm, misogyny, profanity, and other forms of abusive language (Das et al., 2024). Developing automated systems capable of detecting polarization is therefore essential for improving content moderation on digital platforms. However, existing automated moderation tools are predominantly designed and perform substantially better in English and struggle for low-resource languages (Nicholas and Bhatia, 2023). A significant resource gap persists for many non-English languages, limiting the development of robust multilingual moderation systems. SemEval-2026 Task 9: POLAR addresses this gap by introducing a multilingual, multicultural, and multi-event dataset for polarization detection, comprising over 110,000 instances across 22 languages drawn from diverse online platforms and real-world events (Naseem et al., 2026a)[1]. The shared task contains 2 subtasks. Subtask 1 is a binary classification task to determine whether a post contains polarized content. Subtask 2 is a multi-label classification task to classify the type or target polarization for a given text (categories like Political, Racial/Ethnic, Religious, Gender/Sexual or Other). Subtask 3 is also a multi-label classification task to classify how polarization is expressed, with multiple possible labels including Stereotype, Vilification, Dehumanization, Extreme Language, Lack of Empathy, or Invalidation.

In this work, we present our system for multilingual polarization detection for Subtask 1. Our approach leverages transfer learning by fine-tuning a pretrained encoder-only transformer model on auxiliary tasks closely related to polarization, like hate speech detection, implicit abusive language identification, toxicity, offensive speech and gender-based violence (Mozafari et al., 2019). By incorporating task related data, we aim to improve the model's ability to capture nuanced and context-dependent forms of polarized text. In addition, we also evaluate different classification approaches, like zero-shot and few-shot polarization detection using large language models (LLMs), as well as safety-oriented LLMs such as Llama-Guard (Kumar et al., 2024; Saha et al., 2022; Melis et al., 2025; Ranjan et al., 2025). These experiments assess the

[1] https://polar-semeval.github.io/

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2954–2959

effectiveness of general-purpose instruction-tuned models and specialized safety classifiers on the multilingual polarization detection task. Our system focuses on detection for all 22 languages in Subtask 1.

2 Dataset

The dataset is constructed by aggregating multiple resources related to hate speech, toxicity, polarization, and other forms of online harm across different languages. It consists of textual data collected from social media and online platforms, including news websites, Reddit, blogs, Bluesky, and regional forums (Naseem et al., 2026b). The content spans a range of topics such as elections, conflicts, gender rights, and migration. Each language includes approximately 3,000–5,000 annotated instances. Overall, the task covers 22 languages representing diverse cultural and geographical contexts: Amharic, Arabic, Bengali, Burmese, Chinese, English, German, Hausa, Hindi, Italian, Khmer, Nepali, Odia, Persian, Polish, Punjabi, Russian, Spanish, Swahili, Telugu, Turkish, and Urdu. The dataset statistics for Subtask 1 are reported in Table 1.

Statistic	Value
Training Data	73,681
Development Data	3,687
Test Data	33,288

Table 1: Dataset Statistics

3 Methodology

Subtask 1 is formulated as a multilingual binary text classification problem. Given an input text sequence x from a set of languages $\mathcal{L}$, the objective is to predict whether the text contains polarized content. The task is to learn a function $f(x) \rightarrow y$, where $y \in \{0, 1\}$, with 1 indicating that the text is polarized and 0 indicating its not. Model performance is evaluated using F1-macro score.

3.1 Fine-tuning Encoder-Only Models

Looking at prior work (Ragab et al., 2025), encoder-only language models (LMs) have been widely used for natural language understanding tasks, including classification, and have achieved state-of-the-art performance (Marone et al., 2025). Multilingual extensions of these models, such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020), further enable cross-lingual modelling. Recent analyses indicates that encoder-only models are significantly more effective for classification tasks than decoder-only models (e.g., large language models) at comparable parameter sizes, and can outperform decoder models that are an order of magnitude larger (Weller et al., 2025; Gisserot-Boukhlef et al., 2025). To set a baseline we fine-tuned the models for Subtask 1.

The models were selected based on prior work (Plaza-del Arco et al., 2023; Aluru et al., 2020; Naseem et al., 2026b). Specifically, seven well-performing multilingual encoder models were chosen: XLM-RoBERTa-base (Conneau et al., 2020), mBERT-base (Devlin et al., 2019), mDeBERTa-base (He et al., 2021), GTE-multilingual-MLM-base (Zhang et al., 2024), MMBERT-base (Marone et al., 2025), Glot500-base (Imani et al., 2023), and HateBERT (Caselli et al., 2021). All models were fine-tuned using a standard sequence classification setup with a task-specific classification head. Training employed cross-entropy loss with commonly used hyperparameters, like tuned learning rates, batch sizes, number of epochs, validation-based model selection, and early stopping to prevent overfitting.

3.2 Zero and Few-Shot Classification using LLMs

We additionally evaluated the ability of LLMs to detect polarization without task-specific fine-tuning. Due to computational and budget constraints, we focused on open-weight multilingual LLMs. Many publicly available models exhibit strong performance primarily in high-resource languages; therefore, we selected models with broader multilingual coverage. Specifically, we evaluated Qwen2.5-8B, Llama-3.1-8B, and LlamaGuard-3-8B. LlamaGuard was included to assess whether models trained for safety and content moderation tasks generalize to polarization detection. For zero-shot classification, the model was prompted with the input text and asked to determine whether it contained polarized content. For few-shot classification, the prompt additionally included 15 labeled examples of polarized text and 15 examples of non-polarized text sampled from the training set. Predictions were obtained by mapping the model's generated responses to binary labels.

Dataset	Type	Lang	Size
Hate Check HIn	Hate Speech	Hindi	4.75K
HateXplain	Hate/ Offensive Speech	English	20.15K
Implicit Hate	Implicit Hate Speech	English	21.48K
MACD	Explicit Abuse	Hindi	33.64K
OLID	Offensive Speech	English	14.1K
ToxiGen	Toxicity	English	9.91K
Uli	Gender Violence	Hindi	14.78K

Table 2: Dataset Description for Transfer Learning

3.3 Transfer Learning from Related Tasks

Polarized discourse often overlaps with other forms of harmful language, such as hate speech, toxicity, abusive language, and gender-based violence. We explored a transfer learning approach in which a model is first fine-tuned on related tasks before being fine-tuned on the polarization dataset. We compiled multiple datasets covering hate speech detection, implicit abuse, offensive language, toxicity, and gender-based violence across English and Hindi. To keep the focus limited, we decided to only select languages spoken by the authors i.e. English and Hindi. Table 2 summarizes these datasets. The datasets used are HateCheckHIn (Das et al., 2022), HateXplain (Mathew et al., 2021), Implicit Hate (ElSherief et al., 2021), MACD (Gupta et al., 2022), OLID (Davidson et al., 2017), ToxiGen (Hartvigsen et al., 2022) and Uli (Arora et al., 2024).

Based on the baseline experiments, mDeBERTa achieved the best performance among encoder models and was therefore selected as the base model for transfer learning.

The training procedure consisted of sequential fine-tuning: the model was first trained on the aggregated auxiliary datasets and subsequently fine-tuned on the Subtask 1 training data. This approach aims to provide the model with richer representations of harmful discourse, potentially improving its ability to detect nuanced and implicit forms of polarization.

4 Results

Table 3 presents the Macro-F1 scores obtained by the different approaches evaluated in this work. Among the supervised encoder-only models, mDeBERTa achieved the best performance (0.759), outperforming other multilingual encoders such as XLM-RoBERTa and mBERT.

Zero-shot classification using LLMs yielded substantially lower performance compared to supervised fine-tuning, indicating that polarization detection remains challenging without task-specific adaptation. Few-shot prompting improved performance, particularly for Qwen2.5-8B, but still did not match the best supervised models. Safety-oriented LLMs such as LlamaGuard showed limited effectiveness on this task, this means that such models are trained for explicit safety hazards and do not necessarily transfer well to polarization detection which is impliict and contexual in nature.

The transfer learning approach achieved the highest overall performance, with mDeBERTa reaching a Macro-F1 score of 0.781 after pretraining on related harmful language tasks. This result indicates that auxiliary supervision from related domains provides useful signals for detecting polarization, especially for implicit or context-dependent cases. Overall, our findings highlight the effectiveness of supervised multilingual encoder models and demonstrate that transfer learning from related tasks can substantially improve performance, while prompting-based approaches with LLMs remain less competitive in this setting.

5 Conclusion

In this paper, we presented our system for multilingual polarization detection for Subtask 1 of SemEval-2026 Task 9. We explored three modeling paradigms: supervised fine-tuning of multilingual encoder-only models, zero and few-shot classification using large language models, and transfer learning from related harmful language tasks. Our experiments show that supervised encoder-only models provide strong performance for multilingual polarization detection, with mDeBERTa outperforming other evaluated architectures. Prompt-based approaches using LLMs were less effective, particularly in zero-shot settings, suggesting that polarization detection remains challenging without task-specific supervision. The best performance was achieved through sequential transfer learning, demonstrating that knowledge from related

Model	F1-macro
Fine-tuning Encoder-only models	
XLM-RoBERTa	0.693
mBERT	0.660
mDeBERTa	0.759
GTE-multilingual-MLM	0.664
mmBERT	0.708
Glot500	0.670
HateBERT	0.639
Zero Shot Classification	
Qwen-2.5-14B	0.600
Few Shot Classification	
Qwen2.5-8B	0.694
Llama-3.1-8B	0.579
LlamaGuard-3-8B	0.557
Transfer Learning	
mDeBERTa	0.781

Table 3: Results

domains such as hate speech and abusive language can significantly enhance the detection of polarized discourse.

Acknowledgments

We thank Dr. Çağrı Çöltekin for the feedback and proofreading of the work. The authors acknowledge support by the state of Baden-Württemberg through bwHPC for providing compute and GPU access.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Arnav Arora, Maha Jinadoss, Cheshta Arora, Denny George, Haseena Khan, Kirti Rawat, Seema Mathur, and 1 others. 2024. The uli dataset: An exercise in experience led annotation of ogbv. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 212–222.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. 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. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022. HateCheckHIn: Evaluating Hindi hate speech detection models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5378–5387, Marseille, France. European Language Resources Association.

Susmita Das, Arpita Dutta, Kingshuk Roy, Abir Mondal, and Arnab Mukhopadhyay. 2024. A survey on automatic online hate speech detection in low-resource languages. *arXiv preprint arXiv:2411.19017*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hippolyte Gisserot-Boukhlef, Nicolas Boizard, Manuel Faysse, Duarte M Alves, Emmanuel Malherbe, André FT Martins, Céline Hudelot, and Pierre Colombo. 2025. Should we still pretrain encoders with masked language modeling? *arXiv preprint arXiv:2507.00994*.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, and 1 others. 2022. Multilingual abusive comment detection at scale for indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the 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*.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, and 1 others. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

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

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. A modular taxonomy for hate speech definitions and its impact on zero-shot llm classification performance. *arXiv preprint arXiv:2506.18576*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International conference on complex networks and their applications*, pages 928–940. Springer.

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

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

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th workshop on online abuse and harms (woah)*, pages 60–68.

Mohamed Ibrahim Ragab, Ensaf Hussein Mohamed, and Walaa Medhat. 2025. Multilingual propaganda detection: Exploring transformer-based models mbert, xlm-roberta, and mt5. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 75–82.

Rishabh Ranjan, Likhith Ayinala, Mayank Vatsa, and Richa Singh. 2025. Multimodal zero-shot framework for deepfake hate speech detection in low-resource languages. *arXiv preprint arXiv:2506.08372*.

Punyajoy Saha, Divyanshu Sheth, Kushal Kedia, Binny Mathew, and Animesh Mukherjee. 2022. Rationale-guided few-shot classification to detect abusive language. *arXiv preprint arXiv:2211.17046*.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *arXiv preprint arXiv:2507.11412*.

Zecharias Zelalem and Peter Guest. 2021. Why facebook keeps failing in ethiopia. *Rest of World*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

A Language Specific Performance

Language	Encoder-only							Zero Shot	Few shot		Safety
	xlm-roberta-base	m-bert-base	m-deberta-base	m-gte-multilingual-mlm-base	glot-500-base	mm-bert-base	hate-bert	qwen2.5-14B	llama3.1-8B	qwen2.5-8B	llama guard3-8B
Amharic	0.5356	0.4196	0.6927	0.4448	0.482	0.6082	0.4613	0.4825	0.5377	0.528	0.4944
Arabic	0.7751	0.7565	0.8149	0.7681	0.7569	0.78	0.6744	0.7375	0.7243	0.7809	0.661
Bengali	0.818	0.7303	0.8336	0.798	0.7636	0.8459	0.6684	0.7908	0.7372	0.8251	0.7594
German	0.7168	0.6918	0.6967	0.6603	0.679	0.7037	0.679	0.6902	0.6389	0.7547	0.5428
English	0.7208	0.7543	0.7956	0.7101	0.7208	0.7506	0.7506	0.7587	0.7001	0.7491	0.5931
Persian	0.7073	0.7494	0.856	0.6342	0.674	0.7916	0.5085	0.5115	0.6184	0.6508	0.4091
Hausa	0.6301	0.665	0.708	0.708	0.6123	0.6131	0.7945	0.5836	0.1187	0.4647	0.5089
Hindi	0.7562	0.6861	0.8057	0.6482	0.7122	0.6162	0.6642	0.5692	0.5291	0.6545	0.4681
Italian	0.5607	0.562	0.6619	0.5804	0.5843	0.5531	0.5715	0.517	0.4547	0.5599	0.5575
Khmer	0.4747	0.5075	0.535	0.4739	0.4747	0.504	0.504	0.0889	0.1132	0.4917	0.1139
Burmese	0.7402	0.7363	0.8806	0.7677	0.737	0.8009	0.5174	0.608	0.5657	0.7695	0.6019
Nepali	0.8084	0.6768	0.798	0.7874	0.7864	0.8286	0.7199	0.6349	0.8279	0.8095	0.712
Odia	0.6878	0.5373	0.7563	0.41	0.7156	0.6775	0.5807	0.6161	0.5873	0.7783	0.5493
Punjabi	0.6991	0.7172	0.7698	0.6604	0.6511	0.7196	0.6186	0.5386	0.655	0.7778	0.6435
Polish	0.7224	0.6531	0.8103	0.6454	0.6558	0.6779	0.6531	0.7141	0.6271	0.6877	0.6351
Russian	0.7068	0.6767	0.7623	0.6554	0.686	0.7459	0.566	0.7149	0.4729	0.6603	0.5999
Spanish	0.6596	0.6718	0.7073	0.6662	0.6202	0.636	0.6248	0.7043	0.603	0.6591	0.5425
Swahili	0.6625	0.7156	0.7306	0.7115	0.6509	0.7612	0.8165	0.4782	0.6781	0.6371	0.7019
Telugu	0.6946	0.644	0.749	0.6525	0.6606	0.6941	0.6615	0.4161	0.6562	0.5887	0.3699
Turkish	0.7038	0.6515	0.7823	0.7212	0.7116	0.7391	0.6348	0.7718	0.6062	0.7877	0.6174
Urdu	0.6173	0.5192	0.6934	0.6672	0.5873	0.6723	0.6722	0.591	0.602	0.7799	0.5582
Chinese	0.8458	0.7944	0.858	0.836	0.8081	0.8505	0.7103	0.6873	0.6906	0.8831	0.6154

Table 4: Language Performance for all 22 languages