Online Polarization Detection in Persian (Farsi) Social Media

Saeedeh Davoudi
Information Retrieval Lab
Computer Science Department
Georgetown University
saeedeh@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Computer Science Department
Georgetown University
nazli@ir.cs.georgetown.edu

Abstract

Polarization detection in low-resource and mid-resource languages remains a significant challenge for social understanding. This paper presents the first comprehensive benchmark to evaluate transformer-based models for detection of polarized language in Persian (also called Farsi) social media. The aim is to evaluate 1) how and if finetuning the pre-trained models have substantial impact; 2) how Persian specific monolingual models compare to multilingual for this task; 3) how and if transfer learning from models trained on other languages such as culturally-distant English, and culturally-close[er] Turkish, and Arabic can be of interest for this task; and 4) how competitive Large Language Models (LLMs) are in a zero-shot setting. Our evaluation of ten transformer-based models and two LLMs on a publicly available Farsi polarization dataset shows promising findings, highlighting both the strengths and limitations of each approach.

1 Introduction

Polarization is the fragmentation of society into antagonistic groups with fundamentally opposed values and identities (Stewart and Tingley, 2020). Polarization in online discourse has emerged as a challenge for social media platforms and content moderation systems (Shen and Rose, 2019). Polarization is often discussed under the umbrella of toxic or hateful language, but it captures a different phenomenon. Hate speech focuses on how a target is spoken about, while polarization focuses on how sharply groups are separated in public discourse. For example, a polarized sentence in Persian might say:

«کسی که از این حکومت دفاع می‌کند، هیچ‌وقت نمی‌تواند درد مردم معترض را بفهمد؛ دنیای ما کاملاً از هم جداست.»[1]

This sentence clearly separates us from them and presents the groups as unable to understand each other, but they do not directly insult each other. In contrast, a hate-oriented sentence might be:

«معترض‌ها آدم‌های کثیفی هستند و باید از جامعه جمع شوند.»[2]

Here, a group is directly attacked, which turns the statement into hateful or abusive content.

While significant progress has been made in detecting polarization in high-resource languages like English (Juvino Santos et al., 2025), low-resource and mid-resource languages such as Persian remain understudied despite having large, politically active online communities. This gap represents a critical limitation: millions of Farsi speaking users express polarized views on Twitter, Instagram, and other platforms regarding politics, religion, women's rights, and social issues. Yet, tools to systematically detect and analyze this polarization remain largely unavailable. The recent expansion of the POLAR dataset to include Persian (Naseem et al., 2025) presents an opportunity to evaluate polarization detection models for this language.

This paper presents the first comprehensive benchmark for the detection and categorization of Persian polarization in both binary and multilabel settings. We investigate four key research questions. First, to what extent does fine-tuning improve performance compared to pretrained models? This question examines whether domain-specific training substantially enhances model capabilities. Second, how do monolingual (specifically pretrained for Persian) models compare with multilingual models on Persian polarization detection and categorization? Third, does cross-lingual transfer from models trained on languages such as Arabic and Turkish, which are spoken in countries

[1] Translation: "Someone who defends this government can never understand the pain of the protesting people; our worlds are completely separate."

[2] Translation: "The protesters are filthy people and should be cleared away from society."

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 50–59

with some level of cultural similarity, outperform transfer from the culturally distant English language? Fourth, how well are LLMs for both binary polarization detection and multilabel polarization-type classification?

In the rest of this paper, we first review related works in Section 2. Section 3 and Section 4 present the datasets used in our study, the models employed, and the experimental setup and results. Finally, Section 5 outlines the limitations and directions for future work, and Section 6 concludes the paper.

2 Related Work

Polarization detection has emerged as an important research area in Natural Language Processing (NLP), particularly given its implications for understanding online discourse and social dynamics. Early work focused primarily on English-language contexts, examining ideological expression in political debates and social media (Kubin and Von Sikorski, 2021; Hohmann et al., 2023; Kreiss and McGregor, 2024). (Naseem et al., 2025) introduced POLAR, the largest multilingual polarization benchmark with more than 23,000 instances across 7 languages, including multi-level annotations for binary detection, type classification, and manifestation identification. The dataset addresses a critical gap in multilingual NLP by providing balanced coverage of diverse languages and social phenomena related to political, ethnic, religious, and gender-based polarization. The recent expansion of the POLAR framework includes a Persian dataset, offering an unprecedented opportunity to evaluate polarization detection models for low and mid-resource languages.

Persian NLP has seen growing attention, particularly in toxic language detection, hate speech detection, and offensive language identification. A foundational contribution to Persian hate speech research is the PHATE dataset (Delbari et al., 2024), which consists of over 7,000 manually-annotated Persian tweets with multi-label hate speech annotations. What distinguishes PHATE from previous Persian datasets is its inclusion of annotators' explanations that justify each hate speech label alongside information about targeted groups. This enriched annotation approach facilitates not only supervised classification but also research on model interpretability. The dataset encompasses a hierarchical multi-label structure with three subcategories: violence, vulgarity, and hate. (Kebriaei et al., 2023) contributed to Persian offensive language detection research by constructing a large-scale 38,000-tweet corpus using keyword-based data selection techniques combined with crowdsourced annotations and a curated Persian insulting lexicon. This dataset represents one of the largest openly available resources for Persian harmful language detection and demonstrates the feasibility of scaling Persian offensive language annotation. The study employed both classical machine learning approaches and modern transformer-based models to establish baselines on this dataset. Their work shows that Persian-specific models outperform multilingual alternatives, reinforcing the importance of language-specific pre-training for this task.

Cross-lingual transfer has proven effective for many NLP tasks. Recent work by (Bokaei et al., 2025) provides particularly relevant insights for cross-lingual transfer learning in Persian toxic language detection. In their comprehensive study on the PHATE dataset, they investigated the role of cultural context in transfer learning effectiveness. Critically, they found that transfer from languages originating from culturally similar countries (Arabic, Indonesian) yields significantly better results than transfer from culturally distant yet resource-rich languages like English.

However, the focus of prior work on toxic language and hate speech detection has left polarization detection, a distinct phenomenon with different manifestations and social implications, largely unexplored in Persian.

3 Experimental Plan

We structure our experimental plan around the following research questions on Persian polarization detection and polarization categorization:

- **RQ1:** Does fine-tuning on Persian polarization data improve performance compared to using pretrained models alone? Which fine-tuned model outperforms the others among monolingual and multilingual models?
- **RQ2:** How do Persian-specific monolingual models compare with multilingual models?
- **RQ3:** Does cross-lingual transfer from culturally related languages, such as Arabic and Turkish, outperform transfer from culturally distant languages such as English?

Table 1: Dataset statistics for polarization detection and category classification. For Persian, values are averages over 5 CV folds with 80% for training and 20% test (10% of training data is considered as validation dataset). Samples can belong to multiple categories.

		Polarization Detection			Category Classification				
Language	**Split**	**Total**	**Non-Polarized**	**Polarized**	**Political**	**Racial**	**Religious**	**Gender**	**Other**
Persian	Train	2,372	632	1,740	1,043	58	229	142	574
	Validation	264	70	194	116	6	25	16	64
	Test	659	171	488	289	16	63	39	160
English	Train	2,676	1,674	1,002	996	264	106	67	121
Turkish	Train	2,364	1,209	1,155	1,057	400	360	113	114
Arabic	Train	3,379	1,868	1,511	780	583	283	369	565

- **RQ4:** Are zero-shot LLMs performing better than transformer-based models in both binary polarization detection and multilabel polarization-type classification?

Polarization Detection We first examine whether a given text contains polarizing content. Each sample is assigned a binary label indicating the presence (1) or absence (0) of polarization. This setting reflects a common real-world scenario in which systems must distinguish polarized discourse from neutral content before any further analysis.

Polarization Categorization We then study a more fine-grained setting that focuses on identifying the types of polarization expressed in a text. Each sample is evaluated independently across five dimensions: Political, Racial/Ethnic, Religious, Gender/Sexual, and Other. The Other category captures content that does not fall into the predefined categories. Since multiple forms of polarization can co-occur within the same text, this is formulated as a multi-label classification problem. For example, gender-related polarization may overlap with religious arguments, which makes this setting more challenging than binary detection.

Dataset Overview We use POLAR dataset (Naseem et al., 2025) for both polarization detection and categorization tasks. Table 1 presents dataset statistics across Persian, English, Turkish, and Arabic languages. It is shown that the dataset is highly imbalanced among labels, and models need to tackle this challenge. Figure 1 presents the sentence length distribution across different languages. Word counts range from an average of 11 words in English to 22 words in Turkish, while character counts vary more, from 70 characters (English) to 172 characters (Turkish). This difference in character/word count reflects the distinct morphological properties of different languages. Persian exhibits a unique pattern where non-polarized content is longer compared to polarized content. In contrast, English, Turkish, and Arabic show the opposite trend, with polarized content being longer than non-polarized content.

Model Selection We evaluate ten transformer-based models, grouped into monolingual Persian models and multilingual models. This design allows us to directly compare language-specific pretraining with multilingual representations and to analyze their strengths and limitations for Persian polarization analysis. In addition, we include two zero-shot LLM baselines, Gemini-2.5 flash lite (Google Cloud) and GPT-5 nano (OpenAI) to assess how instruction-tuned LLMs perform without task-specific training. Table 2 shows all transformer-based models and their categories used in this study.

Monolingual Persian Models We include four models that are pretrained exclusively on Persian data. These models are expected to capture Persian-specific syntax, vocabulary, and discourse patterns more effectively than multilingual models.

ParsBERT is a BERT-style encoder pretrained on a large and diverse Persian corpus and has shown strong performance across multiple Persian NLP tasks (Farahani et al., 2021). ALBERT-fa is a lightweight alternative based on ALBERT, pretrained on billions of Persian tokens, and is included to study the trade-off between model size and performance (Farahani, 2022). RoBERTa-fa follows the RoBERTa training strategy adapted to Persian, enabling us to examine the impact of more aggressive pretraining compared to standard BERT

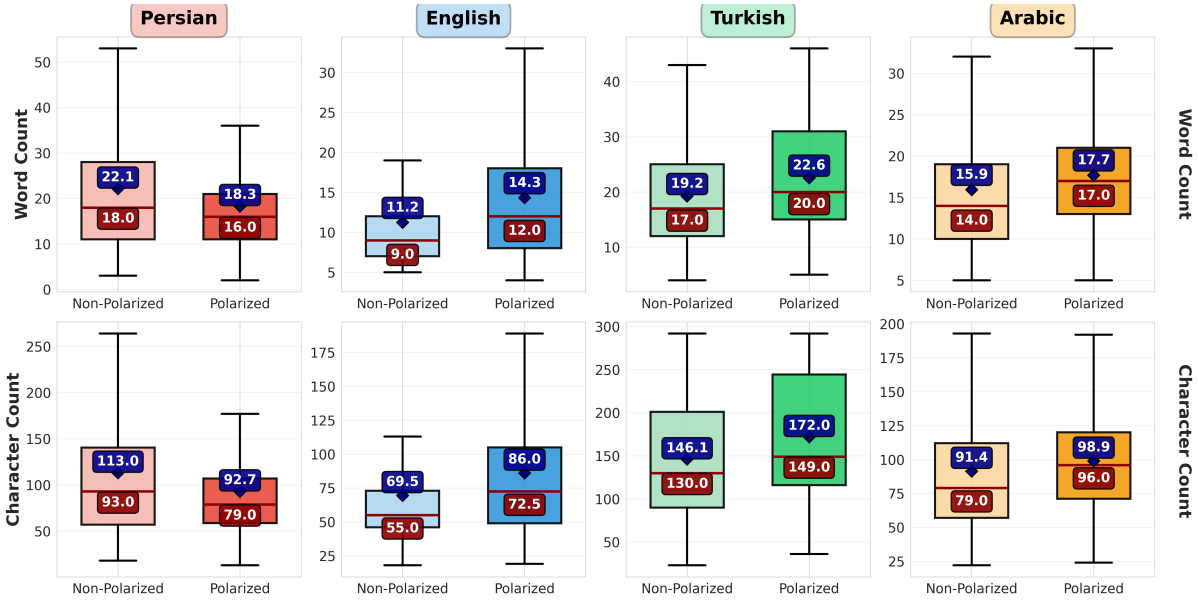

Figure 1: Comparison of sentence length distributions between non-polarized and polarized data across four languages. The top row shows word count distributions, while the bottom row displays character count distributions. Each language is represented with a distinct color scheme: Persian (red), English (blue), Turkish (green), and Arabic (orange), with lighter shades for non-polarized content and darker shades for polarized content. Numbers in small red boxes are median and numbers in small blue boxes are mean.

(HooshvareLab, 2021b). DistilBERT-fa applies knowledge distillation to produce a smaller and faster Persian model while retaining much of the representational power of BERT (HooshvareLab, 2021a).

These models have been widely used and shown strong performance in related classification and representation learning tasks, making them ideal candidates for polarization analysis.

Multilingual Models We also evaluate six multilingual models that differ in architecture, pretraining objectives, and cross-lingual design. These models enable us to assess how well multilingual representations transfer to Persian and how architectural choices affect polarization detection.

mBERT is the original multilingual BERT model trained on Wikipedia data from over 100 languages and serves as a strong baseline for multilingual transfer (Devlin et al., 2019). XLM-R extends mBERT by using larger and more diverse training data and is widely regarded as a strong multilingual encoder (Conneau et al., 2020). InfoXLM builds upon XLM-R by incorporating information-theoretic objectives and parallel data, making it particularly suitable for cross-lingual transfer scenarios (Chi et al., 2021). RemBERT revisits multilingual embedding design and rebalances pretraining data across languages, achieving strong performance on cross-lingual benchmarks (Chung et al., 2020). mDeBERTaV3 enhances cross-lingual generalization through disentangled attention and ELECTRA-style pretraining (Replaced Token Detection), which is particularly effective in low-resource settings (He et al., 2023). Finally, LaBSE is included as a sentence-level encoder trained with translation-ranking objectives, allowing us to study how language-agnostic sentence representations perform compared to token-level encoders (Feng et al., 2022).

Large Language Models Gemini-2.5 flash lite is a cost and latency optimized Gemini 2.5 model designed for high throughput with multimodal input support. GPT-5 nano is the fastest and most cost efficient GPT-5 variant, aimed at high volume workloads like classification and summarization, with text (and image) inputs and text outputs.

Overall, this selection is based on POLAR benchmark and covers a diverse range of model sizes, architectures, and pretraining strategies.

Training Configuration We performed 5-fold crosss-validation with 72-8-20 splits for training, validation, and test. For cross-lingual training, we used a 90-10 split of the source language data for training and validation, then evaluated the trained models on all 5 Persian test folds to assess cross-

Table 2: Overview of models evaluated in this study. The Cross-lingual column indicates if the model was fine-tuned on other languages (English, Turkish, Arabic, and Turkish+Arabic) in addition to Persian. In total, we evaluate 10 pretrained models, 10 finetuned models on Persian, and 24 cross-lingual models.

Model Family	Size	Pretrained Models	Fine-tuned Models on Persian	Cross-lingual
Multilingual Models				
mBERT	178M	mBERT	mBERT-fa	✓
XLM-RoBERTa	278M	XLM-R	XLM-R-fa	✓
RemBERT	575M	RemBERT	RemBERT-fa	✓
mDeBERTa	278M	mDeBERTa	mDeBERTa-fa	✓
InfoXLM	270M	InfoXLM	InfoXLM-fa	✓
LaBSE	471M	LaBSE	LaBSE-fa	✓
Monolingual (Persian) Models				
ParsBERT	118M	ParsBERT	ParsBERT-fa	–
ALBERT	12M	ALBERT-fa	ALBERT-fa-ft	–
DistilBERT	66M	DistilBERT-fa	DistilBERT-fa-ft	–
RoBERTa	125M	RoBERTa-fa	RoBERTa-fa-ft	–

lingual transfer capabilities. It is important to note that samples can belong to multiple polarization categories simultaneously, as reflected in the overlapping category counts shown in Table 1.

All models were fine-tuned with early stopping (patience=5) based on validation performance, with a maximum of 50 training epochs. We used the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 32, and a maximum sequence length of 128 tokens. We also applied class weights during training to address label imbalance. Binary cross-entropy loss was used to handle the multi-label nature of categorization task. [3]

For both binary polarization detection and multi-label categorization, we use the weighted F1 score as the primary evaluation metric, which provides a balanced measure of precision and recall. It accounts for class imbalance by weighting each category's F1 score by its number of true instances, making it more suitable for datasets with varying category frequencies.

4 Results and Analysis

[4] In this section, we present the results and address our research questions. Figure 2 compares pretrained vs finetuned models' performance on two related but distinct tasks: detecting whether content is polarized (binary classification), identifying which specific categories it belongs to (multi-label classification). The evaluation indicates that fine-tuning is essential for both tasks. While pretrained models struggle, fine-tuned models achieve strong results, with the best models reaching 82.7% and 77.9% F1 scores, respectively (RQ1). Also, different models excel at different tasks; category identification is inherently more difficult than polarization detection, with all models showing a consistent performance gap between the two tasks. RoBERTa-fa performs best at detecting polarization (82.7%), while LaBSE-fa leads in identifying specific categories (77.9%) (RQ1). Monolingual models outperform multilingual models in polarization detection. This suggests that for simpler settings such as binary classification, monolingual models are sufficient and well suited to the task (RQ2).

Figure 3 examines whether models trained on polarized content in other languages can detect polarization in Persian without ever seeing Persian training examples. We trained models on English, Arabic, Turkish, and combinations of Arabic and Turkish languages, then tested them on Persian data. The results show that cross-lingual transfer works remarkably well. For binary polarization detection, training on English data produces the best results, with XLM-R achieving 71.5% weighted F1 on Persian test data. This is impressive considering the model never saw any Persian examples dur-

[3] The source code is available at https://github.com/dsaeedeh/Polarization_Detection

[4] When evaluated on the official SemEval 2026 Task 9 blind test set, our fine-tuned system achieved highly competitive results. Specifically, our submission ranked 7th among 44 systems in the first subtask (XLM-RoBERTa model), and 14th among 27 in the second (ParsBERT model).

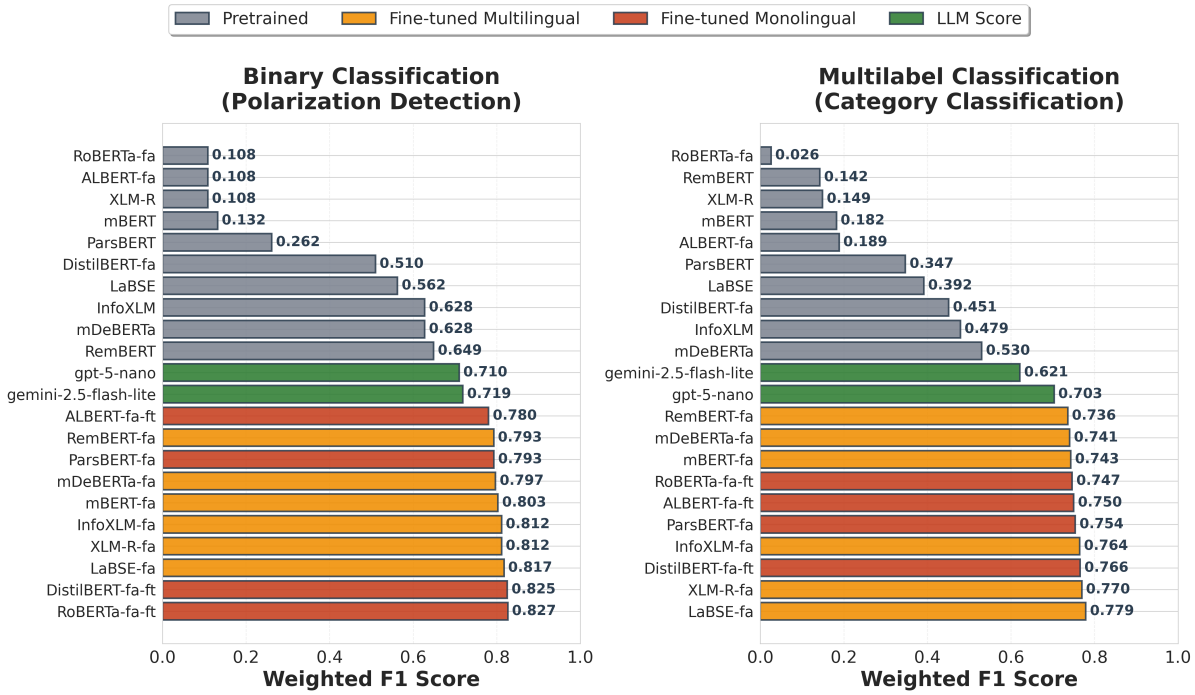

Figure 2: Performance comparison between transformer baselines and zero-shot LLMs for polarization detection (left chart) and categorization (right chart). The plots show that fine-tuning consistently improves performance over pretrained models, while zero-shot LLMs provide strong training-free baselines. Identifying specific categories remains more challenging than detecting whether content is polarized. All bars report 5-fold cross-validation averages, and LLM results are computed by aggregating predictions across folds.

ing training. For multilabel category classification, combining Turkish and Arabic training data yields the best transfer performance at 71.9% weighted F1 using RemBERT (RQ3).

Different models show varying cross-lingual capabilities. XLM-R and RemBERT consistently perform well across different source languages, while other models show more variability. Interestingly, the performance of the "Turkish+Arabic" transfer model likely benefits from cultural proximity. These languages share similar political discourse patterns, religious references, and social issues with Persian, making polarization strategies more transferable across these culturally connected regions (RQ3). These findings show that polarization detection systems can work in new languages without requiring labeled training data in those languages. The 71-72% weighted F1 scores achieved through cross-lingual transfer fall midway between pretrained (37-53%) and fully fine-tuned models (78-83%), offering a practical middle ground when labeled data is scarce.

Figure 4 presents the best performing models for each of the three approaches we evaluated. For binary polarization detection, the results show a clear advantage for fine-tuning. RoBERTa-fa achieves the highest performance at 82.7% when fine-tuned on Persian data. Among pretrained models, RemBERT performs best at 64.9%. Cross-lingual transfer using XLM-R trained on English data achieves 71.5% weighted F1, falling between pretrained and fine-tuned approaches. Turning to multi-label category classification, LaBSE proves to be the superior model, achieving a weighted F1 of 77.9% after fine-tuning. In the pretrained setting, mDeBERTa demonstrates the strongest capability, reaching 53.0%, though a significant gap remains between the fine-tuned and pretrained baselines. Finally, cross-lingual transfer from "Turkish+Arabic" languages using RemBERT yields 71.9%. This pattern can be further understood by examining the cross-lingual category confusion heatmap (see Figure 5). The heatmap reveals that models trained on Turkish and Arabic languages transfer more accurately across culturally salient categories such as Political and Religious, where discourse patterns and topical content are regionally shared with Persian. In contrast, English-trained models exhibit higher confusion, particularly misclassifying Religious, Gender/Sexual, and Political categories. This suggests that the English models lack exposure to the specific cultural and thematic nuances

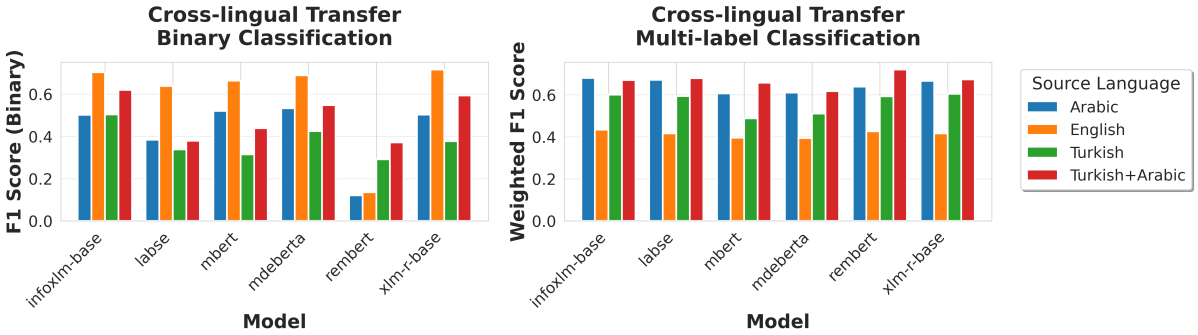

Figure 3: Cross-lingual transfer learning results showing how models trained on other languages perform when tested on Persian. Different colored bars represent training on English, Arabic, Turkish, or Turkish and Arabic data together. For binary polarization detection (left), English training data works best, with XLM-R reaching 71.5% F1. For multi-label categorization (right), Turkish and Arabic combined training achieves 71.9% F1 using RemBERT.

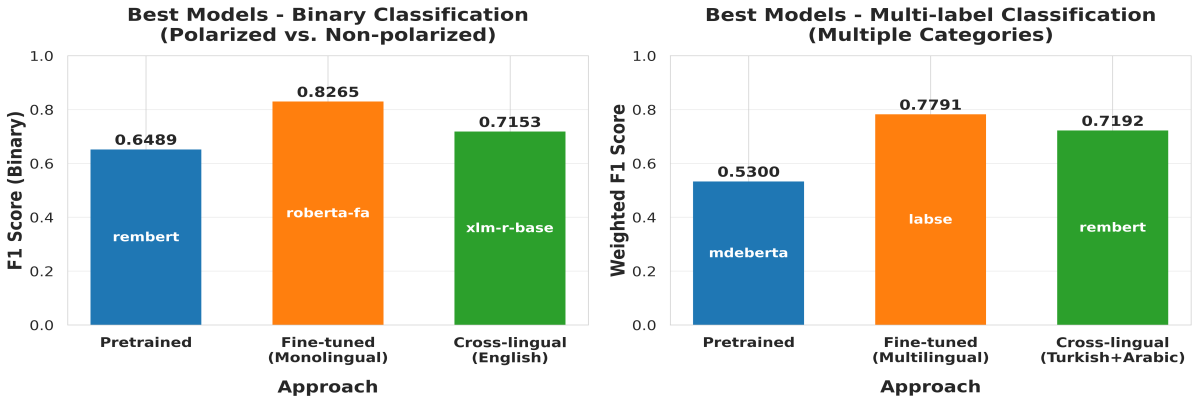

Figure 4: Best performing models for each approach. Each bar shows the top model from three approaches: pretrained, fine-tuned on Persian data, and cross-lingual transfer for binary polarization detection (left) and multi-label category classification (right).

present in Middle Eastern contexts, leading to more frequent misclassifications in categories that are less prominent or differently framed in Western discourse. These findings support the hypothesis that cultural proximity, reflected in shared regional issues and discourse, enhances cross-lingual transfer for fine-grained category classification.

In addition to these transformer-based baselines, we evaluated two zero-shot LLMs, gemini-2.5-flash-lite and gpt-5-nano (Figure 2. For binary polarization detection, gemini-2.5-flash-lite reaches 71.9% weighted F1 and gpt-5-nano reaches 71.0%, placing them close to the cross-lingual transfer result with English-trained XLM-R. This indicates that zero-shot LLMs can provide a strong training-free alternative for polarization detection, outperforming the best pretrained transformer baseline (RemBERT), but still falling notably short of the best fine-tuned model (RoBERTa-fa)(RQ4).

Turning to multi-label category classification, gpt-5-nano achieves 70.3% weighted F1, approaching the culturally proximate transfer result and substantially outperforming the pretrained baseline, while gemini-2.5-flash-lite reaches 62.1%, improving over pretrained transformers but remaining clearly below cross-lingual transfer and fine-tuning models (RQ4).

These results highlight four key findings: First, fine-tuning consistently yields the best performance for both tasks, improving over pretrained models by 27-47% relative. Second, cross-lingual transfer is particularly effective for multi-label classification, closing most of the gap between pretrained and fine-tuned approaches. Third, two zero-shot LLMs provide strong training-free baselines: they outperform the best pretrained transformers, though both remain below fine-tuning. Last, different models excel at different tasks and approaches: RoBERTa-fa dominates binary classification when fine-tuned, while LaBSE leads in multi-label classification, and RemBERT shows strong cross-lingual transfer capabilities.

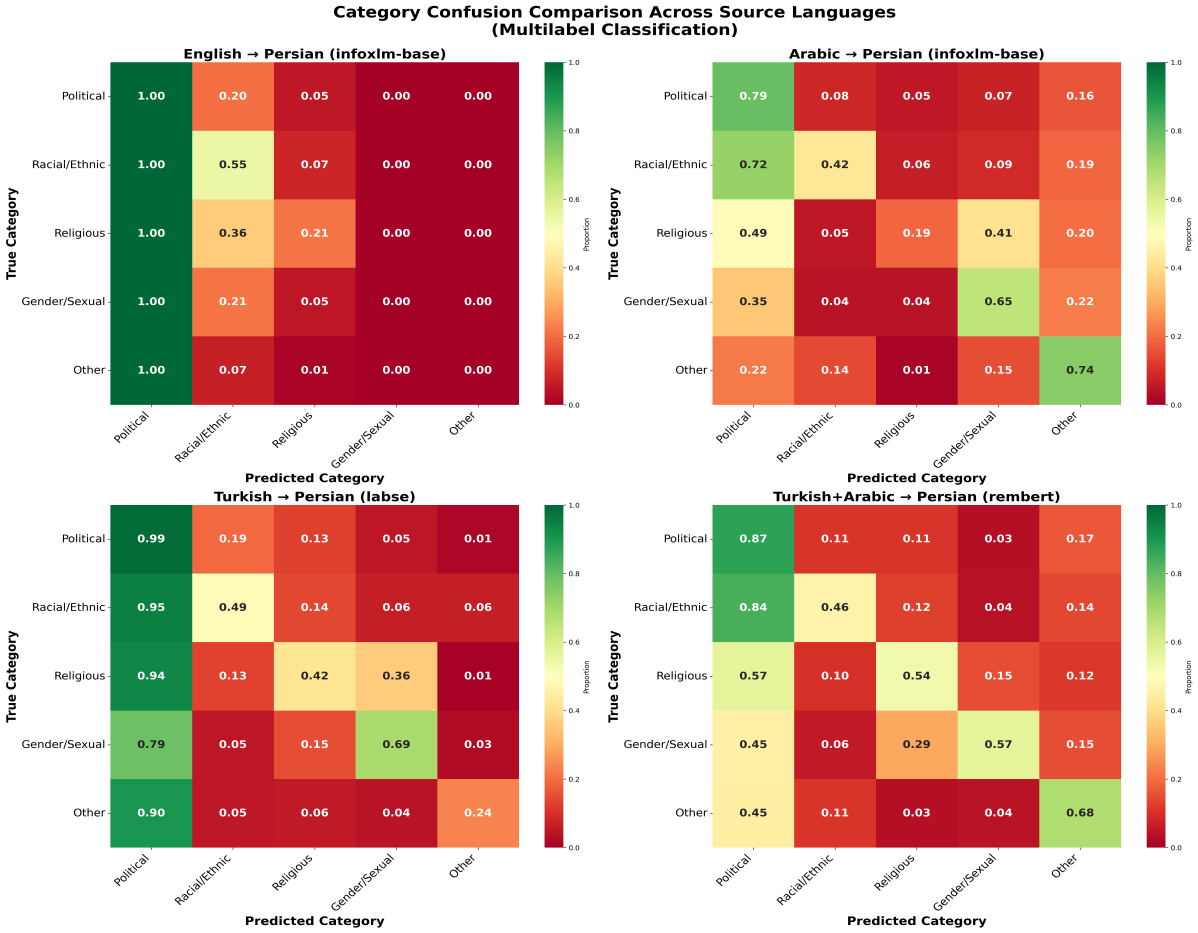

Figure 5: Cross-lingual category confusion matrix comparing best models trained on English, Turkish, Arabic, and Turkish+Arabic. The heatmap highlights that Turkish+Arabic models achieve lower confusion and higher accuracy in culturally salient categories (Political, Religious), while English-trained models more frequently misclassify these categories, especially as Other or Political, reflecting the impact of cultural proximity on fine-grained polarization classification.

5 Limitations

While our study provides comprehensive insights into polarization detection across multiple approaches, several limitations should be acknowledged. Our evaluation focuses on Persian as the primary target language, with cross-lingual experiments on English, Arabic, and Turkish. Expanding to additional languages, particularly those from different language families and cultural contexts, would strengthen claims about cross-lingual transferability. Despite strong gains from fine-tuning, there remains room for improvement, especially for multi-label categorization. Our zero-shot LLM evaluation is limited to two models and a specific prompt setup; performance may vary with prompting strategies, decoding choices, and model versions, and we do not study robustness to such factors. In future work, we plan to extend the LLM setting to few-shot and to develop explainability methods to better understand which linguistic and cultural cues models rely on when detecting polarization across languages.

6 Conclusion

This study benchmarks polarization detection in Persian social media through binary classification (polarized vs. non-polarized) and multi-label categorization (specific polarization types). To compare performance across different learning stages, we tested 10 transformer models under pretrained, fine-tuned, and cross-lingual transfer settings and 2 zero-shot LLMs (gemini-2.5-flash-lite and gpt-5-nano), totaling 46 model instances. Fine-tuning achieves the best results, reaching 82.7% F1 for binary detection and 77.9% weighted F1 for categorization. When labeled data is unavailable, cross-lingual transfer provides a practical alternative, achieving 71-72% F1 and performing particu-

larly well for multi-label categorization when transferring from culturally related languages (Turkish+Arabic). Zero-shot LLMs offer a complementary training-free baseline. They are competitive for binary detection and, for categorization. Overall, these findings show that effective polarization detection is possible in low and mid-resource languages, either through fine-tuning when data is available or through culturally informed cross-lingual transfer and strong zero-shot LLM baselines when data is scarce.

References

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.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

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

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: A persian multi-label hate speech dataset with annotator rationales. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17889–17897.

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.

Mehrdad Farahani. 2022. Albert-persian: A lite bert model for persian. HuggingFace model card m3hrdadfi/albert-fa-base-v2.

Mehrdad Farahani, Mohammad Gharachorloo, and 1 others. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*.

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.

Google Cloud. Gemini 2.5 flash-lite. `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite`. Accessed: 2026-02-10.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Marilena Hohmann, Karel Devriendt, and Michele Coscia. 2023. Quantifying ideological polarization on a network using generalized euclidean distance. *Science Advances*, 9(9):eabq2044.

HooshvareLab. 2021a. Distilbert-fa: A distilled persian bert model. HuggingFace model card HooshvareLab/distilbert-fa-zwnj-base.

HooshvareLab. 2021b. Roberta-fa-zwnj-base: A roberta model for persian language understanding. HuggingFace model card HooshvareLab/roberta-fa-zwnj-base.

Lucas Raniére Juvino Santos, Leandro Balby Marinho, Claudio Elizio Calazans Campelo, Filippo Menczer, and Alessandro Flammini. 2025. Can large language models effectively mitigate polarization in social media text? In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 348–357, New York,

NY, USA. Association for Computing Machinery.

Emad Kebriaei, Ali Homayouni, Roghayeh Faraji, Armita Razavi, Azadeh Shakery, Heshaam Faili, and Yadollah Yaghoobzadeh. 2023. Persian offensive language detection. *Mach. Learn.*, 113(7):4359–4379.

Daniel Kreiss and Shannon C McGregor. 2024. A review and provocation: On polarization and platforms. *New Media & Society*, 26(1):556–579.

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.

Usman Naseem, Juan Ren, Saba Anwar, Sarah Kohail, Rudy Alexandro Garrido Veliz, Robert Geislinger, Aisha Jabr, Idris Abdulmumin, Laiba Qureshi, Aarushi Ajay Borkar, and 1 others. 2025. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

OpenAI. Gpt-5 nano model. `https://developers.openai.com/api/docs/models/gpt-5-nano`. Accessed: 2026-02-10.

Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. Association for Computational Linguistics.

Brandon M Stewart and Dustin H Tingley. 2020. The nature and origins of mass opinion. *Journal of Politics*, 82(3):757–778.