Benchmarking Offensive Language Detection in Persian and Pashto

Zahra Bokaei
School of Informatics
University of Edinburgh
zahra.bokaei@ed.ac.uk

Bonnie Webber
School of Informatics
University of Edinburgh
bonnie.webber@ed.ac.uk

Walid Magdy
School of Informatics
University of Edinburgh
wmagdy@ed.ac.uk

Abstract

Offensive language detection and target identification are essential for maintaining respectful online environments. While these tasks have been widely studied for English, comparatively less attention has been given to other language, including Persian and Pashto, and the effectiveness of recent large language models for these languages remains underexplored. To address this gap, we created a comprehensive benchmark of diverse modeling approaches in Persian and Pashto. Our evaluation covers zero-shot, fine-tuned, and cross-lingual transfer settings, analyzing when detection succeeds or fails across different model approaches. This study provides one of the first systematic analyses of offensive language detection and cross-lingual transfer between these languages.

1 Introduction

With the widespread use of online platforms, offensive language has become a persistent challenge in digital communication, with harmful consequences for individuals and communities (Singh and Li, 2021). Effectively moderating such platforms therefore requires reliable detection of offensive content, as well as identifying its type and intended target to support appropriate responses. Recently, research has increasingly explored offensive language detection beyond English, including work on Iranian languages such as Persian (Ataei et al., 2022; Safayani et al., 2024; Kebriaei et al., 2024) and, to a lesser extent, Pashto (Haq et al., 2023). However, compared to the extensive literature on English, systematic evaluation for these languages remain limited (Ataei et al., 2022), with Pashto in particular receiving substantially less attention.

Persian and Pashto are closely related languages. Persian is spoken primarily in Iran, Afghanistan, and Tajikistan, with approximately 110 million speakers worldwide, while Pashto is spoken mainly in Afghanistan and Pakistan, with an estimated 45–55 million speakers (UNESCO Silk Roads Programme, b,a). Despite their linguistic and cultural proximity, cross-lingual studies of offensive language detection for these languages remain scarce.

Recent advances in large language models (LLMs) have led to substantial gains across many NLP tasks, including harmful content detection. While such models have been widely evaluated for high-resource languages, their effectiveness for Persian and Pashto—particularly for offensive language detection and target identification—remains underexplored (Bokaei et al., 2025). In this paper, we benchmark recent instruction-tuned and multilingual models for offensive language detection in Persian and Pashto across zero-shot, fine-tuned, and cross-lingual transfer settings. We additionally evaluate target identification for individual- and group-directed offenses and analyze cross-lingual transfer between the two languages. Finally, we conduct error analysis to identify systematic model strengths and weaknesses. To the best of our knowledge, this work is the first to jointly study offensive language detection, target identification, and cross-lingual transfer between Persian and Pashto using recent LLMs, with a comprehensive error analysis.

In this study, we address the following research questions (RQs):

RQ1. How do different models perform on offensive language detection in Persian and Pashto?

RQ2. When does Transfer Learning help detect offensive language between Persian and Pashto?

RQ3. How does target type affect offensive language detection in Persian?

To address these research questions, we use two publicly available datasets: Pars-OFF for Persian (Ataei et al., 2022), which contains over 10,000 instances annotated for offensiveness and target type, and POLD for Pashto (Haq et al., 2023), comprising 34,400 instances labeled as offensive or non-offensive. Importantly, target annotations are available only in Pars-OFF, and not in POLD. Our find-

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 13–23

ings reveal consistent patterns across models and languages. *We show that all LLMs reliably detect highly explicit, profanity-based abuse, but systematically struggle with implicit and context-dependent offense.* Fine-tuning substantially improves recall by enabling models to capture language-specific realizations of offensiveness missed in zero-shot settings. Target identification reveals structural limitations, particularly for proxy and mixed targeting, where grammatical form diverges from semantic target. Transfer experiments show a clear asymmetry: Persian-to-Pashto transfer is stronger than the reverse, indicating that cross-lingual effectiveness depends on different factors, such as how offense is encoded in each language.

2 Related Work

This section briefly reviews prior work on offensive language detection and cross-lingual transfer, with a focus on Persian and Pashto. For Persian, multiple datasets benchmark traditional and transformer-based models on social media text (Kebriaei et al., 2024; Safayani et al., 2024; Mozafari et al., 2024), while resources such as Pars-HaO and Pars-OFF provide multi-level and target-aware annotations (Sheykhlan et al., 2023; Ataei et al., 2022). In contrast, Pashto remains severely under-resourced: POLD is currently the only publicly available dataset, and Pashto-specific BERT models outperform multilingual alternatives such as XLM-R on this dataset (Haq et al., 2023). Some prior work explores TL for offensive language detection. Pamungkas and Patti (2019) demonstrates improved robustness through multilingual and cross-domain transfer, El-Alami et al. (2022) shows effective English-to-Arabic transfer with BERT-based models, and Zhou et al. (2023) highlights the limitations of zero-shot transfer due to cultural and contextual mismatch, motivating few-shot and multilingual approaches. Despite this progress, existing work largely focuses on earlier models and does not provide a systematic benchmark of recent LLMs for Persian and Pashto. Moreover, cross-lingual TL between these two closely related Iranian languages with detailed error analysis remain under-explored. Our work addresses these gaps.

3 Dataset

We conduct our experiments on two publicly available datasets for offensive language detection in Persian and Pashto: Pars-OFF (Ataei et al., 2022) for Persian and POLD (Haq et al., 2023) for Pashto. Pars-OFF, a Persian dataset from Twitter data, consists of 10,563 instances, of which 7,381 are labeled as non-offensive and 3,182 as offensive. For the offensive instances, it also includes 2,612 cases targeting individuals, 1,280 targeting groups, and 553 labeled as other targets. For Pashto, POLD Twitter dataset contains 34,400 Twitter instances. Among these, 12,400 instances are labeled as offensive and 22,000 as non-offensive. POLD does not include target annotations.

	Pars-OFF			POLD		
Model	P	R	F1	P	R	F1
GPT-4o	**88**	77	82	**87**	77	82
LLaMA-3-Instruct	83	72	77	78	67	72
Gemma-3-Instruct	74	76	75	75	76	75
Dorna2-Instruct	80	78	79	68	59	60
Mistral-Instruct	70	66	68	71	57	63
Aya-Expanse	69	75	72	59	57	58
LLaMA-3 (FT)	85	**87**	**86**	79	79	79
Gemma-3 (FT)	**88**	82	85	82	**80**	81
ParsBERT	84	82	83	83	72	77
XLM-R	86	84	85	84	82	**83**

Table 1: Offensive language detection performance of LLMs on Pars-OFF and Pashto. FT = fine-tuned

4 Experimental Setup

Motivated by recent benchmark studies demonstrating strong performance of contemporary LLMs on safety- and norm-related tasks (Pourbahman et al., 2025), as well as rapid advances represented by models such as LLaMA 3 and Gemma 3 (Guo and Sarker, 2025), we select a diverse set of recent instruction-tuned and multilingual models for our experiments. We evaluate instruction-tuned LLMs in zero-shot settings: LLaMA-3-Instruct, Gemma-3-Instruct, Mistral-8B-Instruct, Aya-Expanse-8B, GPT-4o, and Dorna2-LLaMA-3-8B. In addition we deployed LLaMA-3-Instruct, Gemma-3-Instruct, ParsBERT (monolingual Persian Model), and XLM-R in fine-tuning and cross-lingual TL settings. Table 5 in the Appendix lists all models used in our benchmarking. For offensive detection on both dataset and target identification on Pars-OFF, we adopt a fixed 80%/10%/10% train/validation/test split across all experiments.

Zero-shot Experiments: We iteratively tested multiple prompts and selected the one with the highest validation performance. The final zero-shot prompt is shown in Figure 1 and 2 in the Appendix.

Fine-tune Experiments: All models were fine-

Model	Group			Individua			Other		
	P	R	F_1	P	R	F_1	P	R	F_1
GPT-40	**83**	81	**82**	81	79	80	58	**75**	**65**
LLaMA 3-Instruct	75	61	67	75	58	65	53	51	52
Gemma 3- Instruct	80	62	70	71	71	71	62	48	54
Dorna 2- Instruct	78	64	70	75	75	75	65	48	55
Llama-3 (FT)	80	**82**	81	71	69	70	59	57	58
Gemma-3 (FT)	**91**	**87**	**89**	82	78	80	**67**	63	**65**
ParsBERT	68	79	73	69	65	67	45	70	55
XLM-R	75	79	77	80	76	78	58	62	60

Table 2: Target identification performance on Persian across LLMs

tuned for 10 epochs on each dataset's training set. Final test results are reported using the checkpoint that achieved the highest validation F1 score.

TL Experiment: LLaMA-3-Instruct, Gemma-3-Instruct, ParsBERT, and XLM-R were fine-tuned for 10 epochs on the source-language training split, selecting the checkpoint with the highest validation F1, and evaluated on the target-language test set.

5 Results

Offensive Detection: Table 1 presents offensive language detection performance across different LLMs. Across both Persian and Pashto, fine-tuned models consistently outperform zero-shot models in F1. On Persian, zero-shot instruction-tuned LLMs achieve F1 scores ranging from 68 to 82, while fine-tuning raises performance to 83–86 F1, with LLaMA-3-Instruct reaching the highest score (F1 = 86). On Pashto, zero-shot performance is notably lower (F1 = 57–72), whereas fine-tuning yields consistent improvements, increasing F1 to 77–83, with XLM-RoBERTa achieving the best result (F1 = 83). These gains are primarily driven by recall improvements: for example, LLaMA-3-Instruct recall increases from 72 to 87 on Persian and from 67 to 79 on Pashto after fine-tuning, while precision remains comparatively stable. In contrast, zero-shot models tend to exhibit higher precision than recall, particularly for Pashto (e.g., GPT-4o precision = 87 vs. recall = 77), indicating conservative predictions that miss offensive instances.

Target Identification: Table 2 reports performance for identifying offensive targets. Performance is highest for group targets across models (F1 = 67–89), followed by individual targets (F1 = 60–80), while other targets are consistently the most challenging (F1 = 52–65). Fine-tuning leads to clear improvements over zero-shot inference for all target types, particularly through recall gains. Gemma-3 fine-tuned achieves the strongest overall performance on group targets (P = 91, R = 87, F1 = 89), whereas GPT-4o obtains the highest F1 on individual targets (P = 75, R = 79, F1 = 80). In contrast, zero-shot models exhibit notably lower recall for individual and other targets indicating difficulty in detecting targets that are implicit or structurally ambiguous.

Model	TL from Pashto			TL from Persian		
	P	R	F_1	P	R	F_1
LLaMA 3	75	86	81	79	89	83
Gemma 3	84	86	85	86	88	87
ParsBERT	57	71	63	82	78	70
XLM-R	79	77	70	79	77	78

Table 3: TL results between Pashto and Persian.

TL Experiments: We evaluate TL between Pashto and Persian using LLaMA-3, Gemma-3, ParsBERT, and XLM-RoBERTa, and compare transfer performance against in-language fine-tuning. The results are shown in Table 3. Overall, TL is more effective from Persian to Pashto than the reverse. When transferring from Pashto to Persian, F1 scores range from 63 to 85: Gemma-3 achieves the best transfer result (F1 = 85), closely matching its in-language Persian performance (F1 = 85), while LLaMA-3 attains F1 = 81, remaining below its Persian in-language score (F1 = 86). In contrast, ParsBERT and XLM-R show larger drops when transferring from Pashto to Persian. Transfer from Persian to Pashto yields stronger gains, with F1 scores between 70 and 87. Gemma-3 again performs best (F1 = 87), exceeding its in-language Pashto result (F1 = 81), while LLaMA-3 also improves under transfer (F1 = 83 vs. 79 in-language). XLM-R maintains stable performance (F1 = 78 vs. 83 in-language), whereas ParsBERT shows limited transferability in both directions (F1 = 63–70). Across models, transfer from Persian to Pashto approaches or surpasses in-language baselines for LLaMA-3 and Gemma-3, while transfer from Pashto to Persian exhibits larger performance drops, particularly for ParsBERT and XLM-R.

6 Discussion

To address RQ1 and RQ2, we analyze offensive language detection from three perspectives: (i)instances detected by all models, (ii) missed by all models, (iii) and detected only after fine-tuning. This highlights which aspects of offensiveness are universally captured, consistently overlooked, or learned through supervision in Persian and Pashto.

6.1 RQ1: Model Performance

Cases detected by all models in both languages: represent a high-signal subset in which offensiveness is realized through explicit lexical cues and direct insult constructions. Across Persian and Pashto, these instances are consistently identified by both zero-shot and fine-tuned models, resulting in near-perfect agreement. This pattern indicates that when offensive intent is overt and targets are explicitly realized, detection is largely architecture- and training-independent. The instances in this category are characterized by extreme lexical explicitness, direct insult speech acts, and clear grammatical realization of targets. They contain dense concentrations of common profanities, sexual and kin-based insults, animalization, and dehumanizing expressions, often stacked within short spans or chained across clauses. Offensive intent is enacted directly rather than implied or reported, and emotional intensity is high. Although some examples reference culturally or politically specific entities, correct detection does not depend on such knowledge, as surface-level abusive markers dominate interpretation. Representative examples are provided in Table 6 in Appendix.

Cases missed by all models: In both languages, these instances lack explicit offensive markers and direct target attachment, relying instead on implicit or discursive forms of hostility. All models converge on non-offensive predictions in such cases, indicating a shared limitation in handling implicit, context-dependent offense. Offensiveness is conveyed implicitly through ideological positioning, moral judgment, evaluative political discourse, slogans, or reflective commentary. These texts frequently resemble legitimate argumentation or social critique, employ neutral or formal vocabulary, and exhibit low emotional arousal. In Pashto, indirect encoding through idioms, proverbs, symbolic group references, and religious or historical framing is particularly prominent. Without explicit surface cues, both zero-shot and fine-tuned models consistently predict non-offensive labels. Examples of these false negatives are reported in Table 7 in Appendix.

Cases detected only after fine-tuning illustrate the contribution of supervised learning. These instances are not reliably captured by zero-shot models but are correctly identified after exposure to in-domain training data. This subset demonstrates that fine-tuning improves sensitivity to non-explicit and discourse-driven expressions of offensiveness. This category includes instances in which offense is implicit, discourse-driven, or culturally embedded rather than lexically explicit. Zero-shot models fail to identify these cases, while fine-tuned models succeed in several cases. The texts express hostility through indirect accusation, moral condemnation, sarcasm, or rhetorical critique, often involving abstract or diffuse targets and limited emotional intensity. Correct predictions after fine-tuning suggest improved alignment with language-specific realizations of implicit offense present in the training data. Illustrative examples are presented in Table 8 in the Appendix. Overall, the results for RQ1 show a consistent pattern across Persian and Pashto: explicit abuse is robustly detected by all models, while implicit offense remains challenging, with fine-tuning partially mitigating this limitation.

6.2 RQ2: TL Effectiveness

We analyze the results of the highest-performing experiment (Gemma 3) for both languages. A clear asymmetry is noticed; transfer from Persian to Pashto is consistently stronger than from Pashto to Persian. In both directions, transfer succeeds for explicit, culturally shared insults; however, Persian-trained models generalize more robustly to Pashto than the other way around. reflecting differences in discourse explicitness and transferable norms. Detailed observation reveals that Persian to Pashto transfer performs particularly well on Pashto offenses that rely on direct profanity, kin-based honor insults, animalization, and moral or religious de-legitimization—patterns that are dominant and highly productive in Persian offensive discourse. Table 9 in the Appendix contains examples that are correctly detected after TL, closely mirroring Persian dehumanization, and moral exclusion. Similarly, politically moralized attacks grounded in shared sociocultural narratives—betrayal, hypocrisy, and foreign dependency —transfer robustly. Religious delegitimization and honor-based insults further strengthen TL, relying on shared Islamic moral frameworks and family-centered notions of shame.

In contrast, Pashto to Persian TL succeeds primarily when Persian offenses adopt Pashto's dominant style of overt, literal abuse with minimal discourse embedding. TL fails when Persian offensiveness is expressed through analytical argumentation, irony, or reflective critique. Pashto to Persian transfer also breaks only when Pashto offense is

implicit, idiomatic, or locally pragmatic rather than lexically explicit. Overall, the observed asymmetry indicates that Persian provides a richer and more generalizable offensive signal for Pashto, while Pashto offers a narrower subset of transferable patterns for Persian. This suggests that TL strength depends not only on how explicitly social norms and offense are expressed, but also on stronger in-language source performance, which enables more transferable supervision

6.3 RQ3: Target Effects

Cases detected by all models exhibit a strong alignment between surface form and semantic target. Offense is explicitly attached to the target through clear grammatical realization: pluralized group nouns, explicit individual reference, or unambiguous syntactic binding between insult and target. In these cases, the addressed and targeted entities coincide, minimizing the need for discourse inference; leading both zero-shot LLMs and FT classifiers converge on the correct target label.

Cases missed by all models show that shared failures arise when there is a mismatch between surface address and semantic target scope. A dominant pattern is proxy or representative targeting, where an individual addressee is used to insult a broader group or institution (e.g., political factions, media organizations, national or gender groups). Models consistently prioritize grammatical address over pragmatic generalization, labeling these cases as individual-targeted even when the semantic intent is clearly collective. Additional failure modes include implicit or abstract targets, metonymic references, proverb-like expressions, and ideological condemnation without explicit pluralization. In such cases, offense is group-directed at the semantic level but realized through individual-directed grammar, leading all models to converge on the same error. Table 1 in the Appendix presents some instances of target identification.

7 Conclusion

We benchmark offensive language detection, target identification, and cross-lingual transfer learning for Persian and Pashto using different LLMs. Results show that performance is driven by how offense is linguistically realized. Across both languages, models reliably detect explicit, profanity-rich abuse but consistently fail on implicit and context-dependent offenses. Fine-tuning improves recall by capturing language-specific realizations missed in zero-shot settings, reducing errors on implicit offense. Target identification further exposes shared structural limitations. Finally, transfer experiments show that Persian-to-Pashto transfer is consistently stronger than the reverse, indicating that effectiveness depends not only on linguistic proximity but also on how explicitly offensive patterns are encoded in the source language and on the overall strength of source-language models.

Limitations

This study has several limitations. First, our experiments are limited to two publicly available Twitter-based datasets—Pars-OFF (Persian) and POLD (Pashto). While standard benchmarks, they may not capture the full diversity of offensive language across platforms, domains, or registers, limiting the generalizability of our findings. Second, Pashto remains severely under-resourced compared to Persian, both in terms of dataset availability and pretrained models. This imbalance likely affects absolute performance and contributes to the observed asymmetry in TL. In particular, the weaker in-language performance for Pashto constrains how much transferable supervision Pashto-trained models can provide for Persian.

Third, our target identification analysis is limited to the target taxonomy provided by Pars-OFF (individual, group, other). The “other” category aggregates heterogeneous and often implicit target types, which may partly explain the consistently lower performance observed for this class. More fine-grained or discourse-aware target annotations could yield additional insights. Future work could focus on providing such resources for more in-depth analysis. While target labels are available for Persian (Pars-OFF), the absence of such annotations for Pashto (POLD) limits direct cross-lingual target analysis. One promising direction is to apply a Persian-trained target identification model to Pashto instances already classified as offensive, generating initial target predictions rather than final labels. These predictions could then be validated by human annotators, reducing annotation effort. Given the stronger Persian-to-Pashto transfer observed for explicit and culturally shared offensive patterns, this approach may achieve sufficient precision on a subset of Pashto data to serve as an effective pre-annotation step.

Ethics Statement

This study analyzes publicly available offensive datasets and does not involve collecting new user data. All datasets were obtained from prior peer-reviewed work or shared tasks that follow established ethical guidelines. Because offensive datasets may contain harmful or toxic language, examples shown in this paper are minimized and presented only when necessary for scientific transparency. No personally identifiable information is included in our datasets or model outputs. All experiments were conducted using anonymized text. Models trained in this work are not intended for deployment without further evaluation, fairness review, and context-specific calibration.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Amin Pourdabiri, Behrouz Minaei-Bidgoli, and Mohammad Taher Pilehvar. 2022. Pars-off: a benchmark for offensive language detection on farsi social media. *IEEE Transactions on Affective Computing*, 14(4):2787–2795.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. *arXiv preprint arXiv:2506.03458*.

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.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The LLaMa 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6048–6056.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: transformer-based model for Persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Yuting Guo and Abeed Sarker. 2025. Benchmarking open-source large language models on healthcare text classification tasks. *arXiv preprint arXiv:2503.15169*.

Ijazul Haq, Weidong Qiu, Jie Guo, and Peng Tang. 2023. Pashto offensive language detection: a benchmark dataset and monolingual Pashto bert. *PeerJ Computer Science*, 9:e1617.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Emad Kebriaei, Ali Homayouni, Roghayeh Faraji, Armita Razavi, Azadeh Shakery, Heshaam Faili, and Yadollah Yaghoobzadeh. 2024. Persian offensive language detection. *Machine Learning*, 113(7):4359–4379.

Marzieh Mozafari, Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Offensive language detection in low resource languages: A use case of Persian language. *PLoS one*, 19(6):e0304166.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 363–370.

PartAI. 2024. Partai/dorna2-llama3.1-8b-instruct. `https://huggingface.co/PartAI/Dorna2-Llama3.1-8B-Instruct`. Accessed: 2025-12-19.

Zahra Pourbahman, Fatemeh Rajabi, Mohammadhossein Sadeghi, Omid Ghahroodi, Somayeh Bakhshaei, Arash Amini, Reza Kazemi, and Mahdieh Soleymani Baghshah. 2025. Elab: Extensive LLMs alignment benchmark in persian language. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 458–470.

Mehran Safayani, Amir Sartipi, Amir Hossein Ahmadi, Parniyan Jalali, Amir Hossein Mansouri, Mohammad

Bisheh-Niasar, and Zahra Pourbahman. 2024. Opsd: an offensive Persian social media dataset and its baseline evaluations. *arXiv preprint arXiv:2404.05540*.

Mohammad Karami Sheykhlan, Jana Shafi, Saeed Kosari, and Saleh Kheiri Abdoljabbar. 2023. Parshao: Hate and offensive language detection on Persian tweets using machine learning and deep learning. *Authorea Preprints*.

Sumer Singh and Sheng Li. 2021. Exploiting auxiliary data for offensive language detection with bidirectional transformers. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 1–5.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

UNESCO Silk Roads Programme. a. Pashto. https://en.unesco.org/silkroad/silk-road-themes/languages-and-endanger-languages/pashto.

UNESCO Silk Roads Programme. b. Persian. https://en.unesco.org/silkroad/silk-road-themes/languages-and-endanger-languages/persian.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. *arXiv preprint arXiv:2303.17927*.

Appendix

[1]

[1]This appendix contains offensive examples used solely for research purposes.

Sentence	Explanation	Gold Label	Predicted Label
Persian: @USER کیر فلا تو کس مادرت … مادر کصه … خارکصه … سسسسسگ ننتو از Damn you, fuck your mother's pussy... pussy mother... sister-fucker... ssssdog... your mother	The insult is directly bound to the target through second-person address and kin-based constructions ("مادرت"، "ننت"). There is no separation between who is spoken to and who is insulted.	individual	individual
خاک تو سرت سایبری… شما بسیجیا اصلا فیلم و سریال میفهمید چیه؟ "Shame on you, cyber troll... Do you Basijis even understand what film and series are?"	The addressee is a single user. But "سایبری" and "بسیجیا" clearly refer to an institutional / ideological group. The speaker uses one person as a stand-in for the group.	Group	Individual
حزبلیای سلطنت‌طلب "Royalist Hezbollah types."	This is a compressed ideological label. No insult word is needed; the condemnation is implicit. The phrase targets a political category, not a person. Without explicit plural insult syntax, models struggle with group target recognition.	Group	Other

Table 4: Target Identification In Persian.

Model	#Params	Reference
LLaMA-3-Instruct	8B	(Dubey et al., 2024)
Dorna2LLaMA-3	8B	(PartAI, 2024)
Gemma-3-Instruct	4B	(Team et al., 2025)
Mistral-Instruct	8B	(Jiang et al., 2023)
Aya-Expanse	8B	(Dang et al., 2024)
ParsBERT	162M	(Farahani et al., 2021)
XLM-RoBERTa	270M	(Conneau et al., 2020)
GPT-4o	–	(Achiam et al., 2023)

Table 5: Models used in this study.

Sentence	Explanation	Gold Label	Predicted Label
Persian: قبر پدر و مادرتون سگ برینه اگر مجاهدین را فالو میکنید خائن های کثیف \ May a dog shit on your parents' graves.	Here we see chained abusive elements: curse construction ("قبر … سگ برینه"), moral condemnation ("خائن"), and profanity ("کثیف"). Targets are explicit and plural, and the emotional intensity is high.	1	1
سید رضی لاشی تر از تو و امثال تو هست؟ مادرسگ کفتار امیدوارم داغ عزیزات یکی یکی رو دلت بمونه بچه لاشی! \ Is Seyyed Razi more of a lowlife than you and people like you? You dog's mother, you hyena — I hope the pain of losing your loved ones one by one stays in your heart. You filthy bastard	This example shows dense insult stacking across clauses: animalization ("کفتار"), kin-based insult ("مادرسگ"), and explicit malediction. The target is directly addressed and repeatedly reinforced. The offense is extreme, emotionally charged, and fully explicit		
Pashto: «شرم نلری غلام خنخیر» Shameless pig's slave.	This is a maximally direct insult with explicit dehumanization ("خنخیر" 'pig') and clear second-person address. The offensive intent is enacted directly, not implied. Emotional intensity is high		
خوان استشهادي ته جنت الفردوس غواړم او تالب خناورو ته جهنم غواړم "I wish the highest paradise (Jannat al-Firdaws) for the young martyr, and I wish hell for the Taliban beasts."	Here we see explicit animalization ("خناورو") paired with extreme moral polarization and direct malediction. Targets are clearly realized as groups, emotional intensity is high, and abusive intent is unambiguous.		

Table 6: Offensive Instances Correctly Predicted by All Models.

Sentence	Explanation	Gold Label	Predicted Label
Persian: «@USER حتما بفرستش، البته حرفت و تفاوت فرهنگی در پوشش و همینطور قوانین پوشش سلطنتی کاملا درسته، ولی بیماری پز دادن و خودنمایی در شبکه‌های اجتماعی هم در بین ما زیاد شده.» Definitely send it. Of course, what you said about differences in clothing culture, as well as royal dress codes, is completely correct—but the obsession with showing off and self-display on social media has also become widespread among us.	This instance is framed as reasonable social commentary and cultural critique. It contains no profanity, no insult lexicon, and no direct target attachment. Any potential offensiveness emerges implicitly through moral judgment (“بیماری پز دادن و خودنمایی”) rather than abusive language. Emotional arousal is low, vocabulary is neutral, and the text resembles legitimate argumentation.	1	0
«امروز داشتم فکر میکردم به این گزاره که: “چقدر بیچاره ست کسی که جرات نکنه به حضرت زهرا بگه؛ مادر!” (منظورم خودمم به خودتون نگیرید)...» Today I was thinking about this statement: ‘How unfortunate someone is who doesn’t dare to call Lady Fatima “mother”’... (I mean myself—don’t take it personally).	This example conveys ideological and religious positioning through reflective commentary rather than attack. While the statement implicitly encodes moral hierarchy and evaluative judgment, it lacks explicit hostility, insults, or target realization. The self-directed disclaimer further reduces perceived aggression.		
Pashto: 😍 ده لوڼې سره کرکټ نه کېږي You can’t play cricket with a headscarf (veil).	This is a Pashto proverb-like construction (“you can’t play cricket with a tail”), which functions metaphorically rather than literally. Any exclusionary or derogatory meaning is *indirect* and culturally encoded, relying on shared idiomatic knowledge. There is no explicit insult, no profanity, and no clear target attachment. Surface vocabulary is neutral, emotional arousal is low, and the emoji further softens the tone.		
د زمري اتحاد له ګیدړ سره.” The alliance of a lion with a jackal	This Pashto expression (“the alliance of the lion with the fox”) uses symbolic animal imagery to convey moral or political judgment indirectly. While it may encode criticism, opportunism, or betrayal, there is no explicit insult marker or target realization.		

Table 7: Offensive Instances Incorrectly Predicted by All Models.

Task:

Determine the type of target of the offense.

Target labels:
• Individual – the offense targets a specific person or clearly identifiable individual.
• Group – the offense targets a group of people defined by nationality, religion, ethnicity, gender, political affiliation, profession, or similar collective identity.
• Other – the offense targets an institution, organization, abstract entity, ideology, media, country, or the target is implicit, symbolic, or not clearly identifiable.

Instructions:
• Base your decision on the semantic intent, not only grammatical form.
• If an individual is used as a representative to insult a broader group, choose Group.
• If the target is abstract, symbolic, or indirect, choose Other.
• Choose the single best label.

Output format:
Answer with exactly one label on a single line:
Individual
Group
Other

Post:
"<TEXT>"

Figure 1: Prompt used for Target Identification

Sentence	Explanation	Gold Label	Predicted Label
Persian: @USER ازین کاسبا بزودی در ایران هم قبل انتخابات خیلی خواهیم دید. بخصوص براندازها و چپهای خریداری شده توسط سپاه برای زدن نامزدهای میانه و اصلاح‌طلب.» We will soon see a lot of these opportunists in Iran as well, especially before the elections—particularly the regime-change activists and leftists bought by the IRGC to attack moderate and reformist candidates.”	This instance expresses indirect accusation and delegitimization through political labeling (“خریداری شده”، “برای زدن”). There is no direct insult word, but the discourse clearly encodes hostility toward a group. Targets are collective and ideologically defined, not directly addressed. Zero-shot models often classify this as neutral political commentary. Supervised models benefit from exposure to in-domain political discourse where accusations of manipulation or treason function as implicit offense.	1	0
« باز قالیباف کار می‌کرد ... شماها که یا آدم می‌کشید و یا عکس یادگاری می‌گیرید، کاری نمی‌کنید که نقد کنیم یه موی قالیباف تو تَن شماها نیست...!» At least Ghalibaf was doing something... You people either kill others or just take souvenir photos—you don’t actually do anything we could even criticize... You don’t have a single hair of Ghalibaf’s worth in you!	This example is offensive in an indirect way. It does not use insults or swear words. Instead, it attacks the target by comparing two sides: one is shown as imperfect but working, while the other is described as morally worse (e.g., killing people or only taking photos). The criticism comes from this comparison, not from explicit abuse. Zero-shot models usually miss this kind of offense because there are no obvious offensive words. Fine-tuned models perform better because they have seen similar contrastive political criticism during training and learn that such comparisons often function as implicit attacks in Persian discourse.		
Pashto: د پوست مادرن نسل په ذهن کې یو لوی مشکل غرب محوره تفکر دی دوی خپله مطالعه نه کوي ټول مالومات له سوشل میډیا انټرن» One major problem in the mindset of the post-modern generation is West-centric thinking. They don’t study on their own; all their information comes from social media and the internet.	This sentence expresses implicit offense through ideological judgment, not through insults. The speaker criticizes a group (“پوستِ مادرن نسل”) by portraying them as intellectually shallow, dependent on Western thinking, and lacking independent study. There is no profanity, no direct insult, and no emotionally charged language. The hostility is conveyed indirectly through delegitimization of values and competence.		
او د یو میلیون تنو شهیدانو په وینو راغلی نظام به د څو تنو د شخصی هوس او ناپوهی ښکار سی And the system that was established with the blood of a million martyrs will fall victim to the personal whims and ignorance of a few people.	This sentence expresses moral condemnation through ideological framing. There is no profanity and no direct insult. The offense comes from accusing a group of betraying a system built on sacrifice. Zero-shot models often miss this because it looks like historical or political commentary. Fine-tuned models can detect it because they learn that this kind of framing is commonly used to delegitimize political actors in Pashto discourse.		

Table 8: Offensive Instances Predicted Correctly Only After Fine-Tuning LLMs.

Task:
Given a social media post written in Persian, determine whether it contains offensive language.

Definition:
Offensive language includes insults, abuse, humiliation, dehumanization, threats, or moral/religious delegitimization directed at an individual or a group.
Offense may be explicit (e.g., profanity, direct insults) or implicit (e.g., indirect attacks, moral condemnation, exclusionary rhetoric).

Instructions:
• Consider both literal wording and implied meaning.
• Do not require profanity for a post to be offensive.
• Criticism, disagreement, or political discussion is NOT offensive unless it attacks, humiliates, or degrades a target.
• If the intent is ambiguous, choose the label that best reflects the overall tone and implied hostility.

Output format:
Answer with exactly one label on a single line:
Offensive
Non-offensive

Post:
"<TEXT>"

Figure 2: Prompt used for Offensive Language Detection

Sentence	Explanation	Gold Label	Predicted Label
Pashto «ستا د خور درکوس غیږ واکاړم «ستا بڼخه وغیم ... ستا ناموس وغیم» I'll rip open your sister's cunt." "I'll fuck your wife... I'll fuck your honor	Family-cantered insults (mother / sister / wife / honor) are highly productive in Persian as well. The insult schema (attacking honor via female kin) is structurally identical across languages. Models trained on Persian already learn this pattern, so they generalize easily to Pashto.	1	1
«مرتده خدای مو تباه کره» «دا ټول د شیطان ابلیس ګروه دي» «خوارج خبیثان کلاب النار» "Heretic, may God destroy you." "They are all the group of Satan, Iblis." "Filthy Kharijites, dogs of hell."	Persian offensive discourse also uses these moral exclusion, and Islamic judgment. Shared Islamic vocabulary and moral framing enable strong transfer. These are explicit moral labels, not subtle pragmatics.	1	1
هره ونه میوه نیسي دغي وني بیا ټول خره نیولي دینه Every tree bears fruit; this tree, though, has only attracted donkey	Meaning is metaphorical and idiomatic, not literal. It requires cultural grounding in Pashto proverb logic. Persian models lack exposure to this implicit evaluative style.	1	0
وروره بعضي چې ګمراه شي بیا لاره په روښانه ورځ هم نه شي موندلی کیدی شي کوم تکلیف به لري او یا پیسي به اخلي په دغه بې ایمانه کار Brother, if some people go astray, they won't be able to find their way even in broad daylight. They might have some trouble or they might get paid for this dishonest work.	No profanity or insult terms appear. The offense is insinuation (corruption, moral failure) expressed as reflective commentary.	1	0

Table 9: TL Performance on Offensive Instances