Cross-Lingual Hate Speech Detection in Low-Resource Languages: The Case of Persian and Kurdish

Shahin Yousefi
Independent Researcher, Iran
sh.yousefi72@gmail.com

Ernesto Luis Estevanell-Valladares
University of Alicante, Spain
University of Havana, Cuba
ernesto.estevanell@ua.es

Ruslan Mitkov
University of Alicante, Spain
ruslan.mitkov@gcloud.ua.es

Abstract

Hate speech detection in low-resource languages remains severely constrained by data scarcity and linguistic complexity. This study presents a comprehensive investigation of Persian-to-Kurdish cross-lingual transfer, evaluating four methodological paradigms: rule-based, traditional machine learning, transformer-based, and large language models, across monolingual and cross-lingual settings. Through a two-phase experimental design, we demonstrate that Persian data can substantially enhance Kurdish detection performance. Notably, models trained solely on Persian data frequently outperform the Kurdish (only), with ParsBERT achieving the strongest overall results. While translation-based data augmentation yields limited gains and introduces noise, model-level cross-lingual transfer proves highly effective. These findings highlight the practical value of leveraging related higher-resource languages for low-resource hate speech detection and provide actionable insights for building more equitable content moderation systems in underrepresented linguistic communities.

1 Introduction

The rapid expansion of social media has amplified both the benefits and perils of online communication. Among these perils, hate speech stands out as a particularly corrosive force, capable of inciting violence, deepening societal divisions, and marginalising vulnerable communities (Poletto et al., 2021). In multilingual and geopolitically complex regions such as Iran and Kurdistan, where Persian (Farsi) and Kurdish serve as dominant languages for millions of users, the challenge is compounded by linguistic nuances, dialectal variations, and script differences. Despite the colours of local context, automated hate speech detection systems for these languages remain underdeveloped, largely due to the scarcity of annotated resources and the inherent difficulties of low-resource NLP settings (Das et al., 2024; Poletto et al., 2021).

This scarcity is particularly acute for Persian and Kurdish, both Indo-Iranian languages that share historical roots yet face distinct resource constraints. Persian benefits from a growing but still limited set of datasets and models (Mozafari et al., 2024; Kebriaei et al., 2024), while Kurdish varieties, especially Sorani, suffer from even fewer annotated corpora and pretrained resources (Saeed et al., 2022). Surveys of hate speech detection consistently highlight that while high-resource languages like English have seen substantial progress through large-scale benchmarks and sophisticated models, low-resource settings continue to rely on fragmented efforts, often with modest performance gains from traditional approaches (Das et al., 2024; Poletto et al., 2021). In the case of Persian, recent efforts have introduced datasets focussed on offensive and hateful content from social media. Yet, challenges such as cultural specificity, code-switching, and evolving online slang persist (Mozafari et al., 2024; Bokaei et al., 2026). For Kurdish, early studies have collected Facebook comments and applied basic classifiers, but systematic evaluations remain rare (Saeed et al., 2022).

Cross-lingual transfer learning offers a promising pathway to address these gaps, particularly for closely related language pairs like Persian and Kurdish (Jiang and Zubiaga, 2026). Recent comprehensive reviews emphasise the transformative role of transformer-based architectures in advancing hate speech detection, while underscoring the persis-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 350–360
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_036

tent challenges in low-resource and multilingual contexts (Das et al., 2024; Ramos et al., 2024). By leveraging model-level adaptations (such as multilingual transformers) and data-level strategies (including dataset merging and translation), it becomes possible to enhance hate speech detection in the lower-resourced language. This study adopts a two-phase experimental design. First, a separate rule-based approach is evaluated as a simple pattern-matching baseline. Phase I then establishes performance across traditional ML, transformer-based, and LLM paradigms. Phase II examines whether augmenting the data with original Persian data and its Kurdish translation meaningfully improves hate speech detection performance on Kurdish. In doing so, it contributes to the broader understanding of how transfer techniques can support more equitable and effective content moderation tools in underrepresented linguistic communities (Das et al., 2024; Bokaei et al., 2026; Saeed et al., 2022).

1.1 Motivation

Persian and Kurdish, as closely related Indo-Iranian languages, offer a valuable opportunity for cross-lingual transfer in hate speech detection. While Persian has benefited from a growing number of datasets and models, Kurdish varieties, particularly Central Kurdish (Sorani), remain severely low-resourced, with limited annotated corpora and pre-trained models available (Ahmadi et al., 2024; Das et al., 2024). This resource imbalance is despite the fact that both languages are widely used in online discourse across Iran and Kurdistan, where hate speech frequently draws on shared cultural and political contexts. Previous research has shown that cross-lingual transfer learning can effectively improve hate speech detection in low-resource settings by leveraging data and models from related higher-resource languages (Bigoulaeva et al., 2021). Motivated by this, the present study investigates whether augmenting Kurdish data with Persian resources, through both model-level (e.g., multilingual and language-specific transformers) and data-level strategies (dataset combination and machine translation), can yield meaningful performance gains, while establishing strong baselines using, traditional ML, and LLM approaches.

This study contributes to low-resource hate speech detection by presenting a comprehensive evaluation across rule-based, traditional ML, transformer, and LLM paradigms; systematically comparing model- and data-level cross-lingual transfer strategies for Persian-Kurdish; and demonstrating the benefits of integrating multiple Persian datasets to address Kurdish data scarcity.

2 Related Work

This section reviews existing research on hate speech detection in low-resource languages, cross-lingual transfer learning, and various detection methods, providing the foundation for the present study on Persian and Kurdish.

2.1 Low-Resource Languages in Hate Speech Detection

Low-resource languages, lack sufficient digital resources, such as corpora, annotated datasets, and linguistic tools, for effective NLP model training, despite often having large speaker populations. Pakray et al. (2025) further elaborate on the challenges and opportunities in developing NLP applications for these languages, emphasising the critical gap between speaker population and available computational resources. Hate speech detection in low-resource languages remains a significant challenge due to limited annotated datasets, linguistic diversity, and cultural specificities. Das et al. (2024) provide a dedicated survey on automatic online hate speech detection in low-resource languages, highlighting the rapid growth of toxic content on social media and the persistent gaps in resources for non-English languages. Similarly, Sharma et al. (2025) survey hate speech detection research across South Asian languages, documenting tasks, datasets, and methods while emphasising the under-representation of many Indo-Aryan and Iranic varieties.

Poletto et al. (2021) offer a systematic review of resources and benchmark corpora for hate speech detection, revealing a heavy bias toward high-resource languages and calling for more diverse datasets. In the context of Persian, Mozafari et al. (2024) explore offensive language detection as a use case for this low-resource language, while Sheykhlan et al. (2023) investigate Pars-HaO using Machine Learning and deep learning on Persian tweets. Kebriaei et al. (2024) further advance Persian offensive language detection through detailed modelling.

For Kurdish, Badawi (2024) applies deep learning to cyberbullying detection, and Ahmadi et al.

(2024) develop language and speech technology resources for Central Kurdish varieties, including corpora from transcribed media as an alternative to traditional fieldwork. Roberts (2024) addresses automated hate speech detection in a low-resource environment, while Gutha et al. (2023) tackle multilingual hate speech and offensive language detection specifically for low-resource languages. Bokaei et al. (2026) benchmark offensive language detection in Persian and Pashto, evaluating zero-shot, fine-tuned, and cross-lingual settings. Fetahi et al. (2025) enhance social media hate speech detection in low-resource languages using transformers and explainable AI. Mahmud et al. (2023) review cyberbullying detection for low-resource languages and dialects. Nkemelu et al. (2023) emphasise the role of context experts in tackling hate speech in low-resource settings like Myanmar.

2.2 Cross-lingual Transfer Learning

Transfer learning and cross-lingual approaches have become central to addressing data scarcity in hate speech detection. Ruder et al. (2019) provide a foundational tutorial on transfer learning in Natural Language Processing, outlining methods that leverage pre-trained models across tasks and languages. It is shown that cross-lingual transfer learning for hate speech detection can be achieved by training on high-resource data (English) and adapting to low-resource targets (German) using bilingual embeddings and data bootstrapping (Bigoulaeva et al., 2021). Jiang and Zubiaga (2026) present a systematic review of cross-lingual offensive language detection, analysing datasets, approaches, and challenges across instance, feature, and parameter transfer strategies. Researchers in (Hämmerl et al., 2024) survey understanding of cross-lingual alignment, while Pallucchini et al. (2026) examine methods for cross-lingual alignment of contextualised representations. Al-Badani et al. (2025) discuss transfer learning and cross-linguistic generalisation in multilingual hate speech detection, reviewing approaches and challenges.

Practical applications in (Firmino et al., 2024) improve hate speech detection using cross-lingual learning from English and Italian to Portuguese, achieving state-of-the-art results. Shi et al. (2022) propose cross-lingual offensive speech identification with transfer learning and adversarial methods for low-resource languages. de Oliveira et al. (2023) use a multilingual approach in cross-lingual transfer learning for hate speech detection. Usman et al. (2025) explore multilingual hate speech detection using translation-based approaches with Large Language Models. Research in (Mollanorozy et al., 2023) specifically investigates cross-lingual transfer learning with Persian as a source language, revealing task-dependent benefits. Narula and Chaudhary (2024) provide a comprehensive review of hate speech detection for multilingual data.

2.3 Hate Speech Detection Methods and Resources

The evolution of hate speech detection methods spans rule-based, traditional ML, transformers, and Large Language Models. Albladi et al. (2025) offer a comprehensive review of hate speech detection using Large Language Models. Ramos et al. (2024) survey automatic hate speech detection in the age of transformers. Lee and Ram (2024) propose a personality-based deep learning approach for hate speech detection.

For Persian and related languages, work (Nasiri and Analoui, 2022) apply transfer learning and data augmentation to Persian stance detection, with implications for hate speech tasks. Pakray et al. (2025) discuss Natural Language Processing applications for low-resource languages, while a survey by Joshi et al. (2025) reviews the development of LLMs for European languages, highlighting datasets and resource challenges in low-resource regional dialects. Ali and Pyysalo (2024) survey Large Language Models for European languages, providing a broader context for multilingual capabilities.

These works collectively underscore the shift toward multilingual models, the value of language-specific resources, and the need for hybrid approaches combining classical methods with modern transformers and LLMs. The current study builds upon this foundation by conducting a three-way comparative evaluation focused on Persian and Kurdish hate speech detection.

3 Data and Preprocessing

This section presents the datasets employed in the study, describes the standardisation procedures applied to ensure cross-lingual compatibility, and details the preprocessing pipeline used for all experiments.

Dataset	Original Size	Cleaned Size
Kurdish	6,882	5,738
Naseza	5,719	5,162
PHATE	8,727	6,999
PHICAD	124,100	99,318

Table 1: Data statistic

3.1 Datasets Description

This study utilises four publicly available datasets to support hate speech detection research in Persian and Kurdish. These datasets were selected for their complementary characteristics in terms of size, annotation quality, and domain coverage.

The *Kurdish Hate Speech Dataset (KHSD)* (Saeed et al., 2022) consists of Central Kurdish (Sorani) texts written in the Arabic-based script. No Latin transliteration was applied; only standard Arabic/Persian character normalisation was performed as part of the unified pre-processing pipeline. This dataset contains 6,882 manually annotated Facebook comments in a binary hate/non-hate scheme. It represents one of the few available resources for the Kurdish language. For Persian, three datasets were employed: *Naseza* (Amirivojdan, 2025), a substantial Telegram-based dataset of 5,719 entries focused on political and social discussions; *PHATE* (Delbari et al., 2024), containing 8,727 tweets with multi-label hate annotations and valuable annotator rationales; and *PHICAD* (Davardoust et al., 2024), a large-scale collection of more than 124,100 Instagram comments labelled for harmful content. Table 1 summarises the statistics of the original datasets used in this study.

The combination of these datasets allows for robust evaluation in Phase I (language-specific baselines) (see section 5.2). It enables data-level cross-lingual experiments in Phase II (see section 5.3) by merging Persian and Kurdish resources.

3.2 Data Standardisation

To ensure consistency across all datasets for both language-specific and cross-lingual experiments, a unified format consisting of three columns, *id* (starting from 0), *text*, and *label* (1 for hate, 0 for not hate), was adopted. For the **Kurdish Hate Speech Dataset (KHSD)**, an *id* column was added, and the original Class labels were mapped to binary values (hate $\rightarrow$ 1, not hate $\rightarrow$ 0). The **Naseza** dataset required only remapping of its *label* column to the 0/1 scheme while resetting the *id*. In the **PHATE** dataset, the *HateSpeech* column was used as the binary label, the *text* column was retained, and a new sequential *id* was assigned. For **PHICAD**, rows labelled as *spam* or *spamobscene* were removed, the *comment_normalised* column was selected as *text*, and the *class* column was mapped to binary labels (hate and hateobscene $\rightarrow$ 1; obscene and clean $\rightarrow$ 0). This standardisation process enabled seamless integration of Persian and Kurdish data in the subsequent experimental phases.

3.3 Preprocessing Pipeline

All datasets underwent a uniform preprocessing pipeline. First, empty rows and exact duplicate texts were removed to prevent data leakage. Texts shorter than three words were discarded to retain only samples with sufficient semantic context. Subsequently, emojis were removed, repeated spaces collapsed, and Persian/Arabic character variants were normalised to a standard representation. Diacritics were stripped, newline and carriage return characters removed, and common social media noise, including URLs, user mentions, hashtags, excessive punctuation, textual emoticons, and numeric digits, was systematically cleaned. This deterministic pipeline was applied identically across all Persian and Kurdish datasets before splitting and model training, transforming raw social media content into clean, standardised texts while preserving original labels (0 for not hate, 1 for hate).

4 Methodology

This study adopts a multi-paradigm approach to hate speech detection, comparing four distinct methodological frameworks across Persian and Kurdish datasets. The experiments are structured into two phases: Phase I (see section 5.2) evaluates each language independently on smaller, language-specific datasets, while Phase II (see section 5.3) examines the impact of cross-lingual data augmentation by incorporating Persian data (raw or translated) into Kurdish training sets.

4.1 Paradigm A: Rule-based Approach

Rule-based methods serve as baseline systems that rely on linguistic patterns without requiring training data. These approaches include keyword matching using curated lists of hate-related terms in Persian and Kurdish, pattern matching through regular expressions to capture common hate speech structures (e.g., slurs combined with specific targets),

phrase matching for multi-word expressions, and basic negation handling to reduce false positives. While simple and interpretable, these methods are expected to provide lower performance in morphologically rich and context-dependent languages like Persian and Kurdish.

4.2 Paradigm B: Traditional ML

Traditional Machine Learning models were implemented using TF-IDF vectorisation for feature extraction. Four classifiers were evaluated: Logistic Regression, Linear Support Vector Machine (LinearSVC), NuSVC, and Random Forest. These models were trained on the same standardised datasets as other paradigms. Hyperparameters were tuned using grid search with 5-fold cross-validation on the validation set, optimising for macro-F1 score. This paradigm provides a strong classical baseline against which the benefits of deep learning and transfer learning can be measured.

4.3 Paradigm C: Transformer-based Models

Five transformer models were fine-tuned to capture contextual representations, which are particularly suited for low-resource and cross-lingual settings. These include the multilingual models mBERT (Wu and Dredze, 2020) `bert-base-multilingual-cased`, XLM-RoBERTa (Conneau et al., 2020) `xlm-roberta-base`, and mDeBERTa-v3 (He et al., 2023) `mdeberta-v3-base`, as well as language-specific models: ParsBERT (Farahani et al., 2021) `bert-base-parsbert-uncased` for Persian and KuBERT (Awlla et al., 2025) `KuBERT-Central-Kurdish-BERT-Model` for Central Kurdish. All models were fine-tuned with a maximum sequence length of 128 tokens, a learning rate of 2e-5, a batch size of 16 for training and 32 for evaluation, a weight decay of 0.01, and class-weighted cross-entropy loss to address label imbalance. Training ran for up to five epochs with early stopping based on validation macro-F1. The best checkpoint was selected according to validation performance. This setup allows for a direct comparison of general multilingual capabilities against language-specific pre-training.

4.4 Paradigm D: Large Language Models (LLMs)

Large Language Models were evaluated in zero-shot and few-shot settings without any fine-tuning. The models included `Llama-3.1-8B-Instruct` (Team, 2024), `Qwen3-8B` and `Qwen3-32B` (Team, 2025b), and `Gemma-12B-Instruct` (Team, 2025a). For few-shot prompting, eight balanced examples (four hate, four not hate) were randomly selected from the training set and included in the prompt for each test instance. Prompts were designed to be concise and consistent across languages. Greedy decoding was used to ensure deterministic outputs. Open-weight models were employed; this paradigm assesses the out-of-the-box cross-lingual understanding and reasoning capabilities of modern LLMs in low-resource hate speech detection.

All experiments used the same data splits and preprocessing pipeline to ensure a fair comparison. Evaluation was performed on fixed test sets using Accuracy, macro-F1, per-class F1, and confusion matrices, with macro-F1 serving as the primary metric.

5 Experimental Setup

This section describes the experimental configuration across all four paradigms. It covers the rule-based setup, data splitting for Phase I and Phase II, and the evaluation metrics used to compare model-level versus data-level transfer.

5.1 Rule-based Approach

Paradigm A serves as the non-learning baseline for the study. A unified rule-based system was developed and applied consistently across all datasets without any dataset-specific tuning. The system combines four components: curated keyword lists (hate-related terms in Persian and Kurdish), regular expression patterns to capture common hate structures, phrase matching for multi-word expressions, and basic negation handling to mitigate false positives.

A fixed detection threshold of 1.0 was used, meaning a single match against any rule triggers a hate prediction. This conservative setting prioritises precision over recall and ensures full reproducibility, as the approach is entirely deterministic with no randomness or training involved. The results reveal significant variation across languages

and datasets. The system performed markedly better on Persian datasets (especially Naseza) than on Kurdish, highlighting the challenges of applying simple pattern matching to morphologically complex and lower-resource languages like Kurdish. Detailed confusion matrix analysis confirmed a strong bias towards predicting the non-hate class, particularly in the Kurdish dataset, where hate recall was only 0.216.

5.2 Phase I: Resource-Specific Pipelines and Evaluation

Phase I aims to establish strong language-specific baselines for each language and dataset before investigating cross-lingual effects. This phase evaluates all paradigms (Traditional ML, Transformers, and LLMs) independently on three smaller but representative datasets: the original Kurdish dataset (Kurdish), and two Persian datasets (Naseza and PHATE).

For each dataset, fixed train-validation-test splits were used consistently across all models to ensure fair and reproducible comparisons. The Kurdish test set contains 1,148 samples, while the Persian test sets include 1,033 samples for Naseza and 1,400 samples for PHATE. These splits maintain the original class distribution of each dataset, with hate speech comprising approximately 56% of the Kurdish test set and around 43–45% of the Persian test sets. In this phase, models are trained and tested within the same language and domain. This allows us to understand the inherent difficulty of hate speech detection for each resource and provides a clear reference point against which the gains (or losses) from cross-lingual approaches in Phase II can be measured. All preprocessing steps, including emoji removal, Persian/Arabic script normalisation, diacritic stripping, and whitespace collapsing, were applied uniformly.

5.3 Phase II: Cross-Lingual Transfer and Data-vs-Model Level Enhancement

This section addresses the core research question of this study: to *what extent can Persian data enhance hate speech detection performance in Kurdish through cross-lingual transfer?* This phase systematically compares model-level transfer (using multilingual models on raw data) with data-level transfer (augmenting the training data with Persian content, either in its original or translated form). Only the training data composition differs across the configurations described below.

Three primary training configurations were examined:

- **Persian-only** (Combined_Persian): Models trained solely on the combined Persian datasets (Naseza + PHATE + PHICAD) to establish a pure cross-lingual baseline.
- **Raw cross-lingual transfer** (Combined_Kurdish_Persian): Persian data added directly (without translation) to the original Kurdish training set.
- **Translation-based data augmentation** (MT_Kurdish_Aug): Persian data first translated into Central Kurdish using Google Translate and then combined with the original Kurdish data.

To disentangle the contribution of different components, three additional setups were evaluated on the identical Kurdish test set: **Persian-only cross-lingual** (Persian_xling), **translated Persian only** (MT_Kurdish), and the **fully mixed translated** set (Combined_KP_MT). The Persian-to-Kurdish translation was performed using the Google Translate API. After automatic translation, all outputs were manually reviewed and edited by a native Kurdish speaker to ensure linguistic accuracy and cultural appropriateness. This human post-editing step significantly improved the quality of the augmented data.

Special attention was paid to data splits in Phase II. The validation and test sets were constructed as the union of Phase I validation and test sets, with any overlapping instances carefully removed from the training data.

This design eliminates data leakage and enables a clean, apples-to-apples comparison between Phase I (language-specific) and Phase II (cross-lingual) results. By comparing these configurations, the experimental setup allows us to distinguish between model-level transfer (leveraging multilingual model capabilities on raw data) and data-level transfer (augmenting training data via translation), directly supporting the three-way analysis (Baselines vs Model-level vs Data-level Transfer) of the study.

6 Results and Discussion

This section presents the experimental results from both Phase I (language-specific baselines) and Phase II (cross-lingual transfer), followed by a detailed analysis and comparison across all paradigms.

Dataset	Model	Best macro-F1
Kurdish	LinearSVC	0.609
Naseza	LinearSVC	0.801
PHATE	LogisticRegression	0.699

Table 2: Best macro-F1 scores for traditional ML TF-IDF-based models (Paradigm B - Phase I)

Dataset	Model	Best macro-F1
Kurdish	ParsBERT	0.630
Naseza	ParsBERT	0.893
PHATE	ParsBERT	0.791

Table 3: Best macro-F1 scores for Transformer-based models (Paradigm C - Phase I)

6.1 Evaluation of Rule-based Approach

The rule-based approach (Paradigm A), which relied on keyword lists, regex patterns, phrase matching, and basic negation handling, achieved a macro-F1 of 0.495 on Kurdish, 0.797 on Naseza, and 0.629 on PHATE (see Figure 1). This highlights the challenges of pattern-based methods in morphologically rich languages like Kurdish, where recall for the hate class was particularly low (0.216).

6.2 Evaluation of Language-Specific Baseline (Phase I)

Phase I established strong language-specific baselines across the three datasets. As shown in Figure 2, performance varied significantly by language and dataset.

Traditional ML Models (Paradigm B) using TF-IDF features showed clear improvements over the rule-based baseline. LinearSVC consistently performed best, reaching macro-F1 scores of 0.609 on Kurdish, 0.801 on Naseza, and 0.699 with Logistic Regression on PHATE (see table 2). These results confirm that learning-based shallow models provide substantial gains, especially on the more challenging Kurdish (+0.114) and PHATE (+0.070) datasets, while offering only marginal improvement on Naseza, where the rule-based system was already strong.

Transformer-based models (Paradigm C) further advanced performance. ParsBERT emerged as the strongest model overall in Phase I, achieving 0.630 on Kurdish, 0.893 on Naseza, and 0.791 on PHATE. XLM-RoBERTa and mBERT also delivered competitive results, while mDeBERTa-v3 struggled with training stability on these datasets. Language-specific models (ParsBERT and KuBERT) generally outperformed multilingual ones on their respective languages, though the gap was smaller than expected (see Table 3).

In the LLM paradigm (Paradigm D), few-shot prompting with models like Gemma and Qwen3 showed promising results, particularly on Persian datasets (see Table 4). However, zero-shot performance was generally lower, and all LLMs trailled the best fine-tuned transformers.

Overall, Phase I demonstrates that Persian hate speech detection is relatively more mature, while Kurdish remains challenging even for advanced models.

6.3 Impact of Cross-lingual Transfer Strategies (Phase II)

Phase II directly tested the central hypothesis: whether adding Persian data enhances Kurdish hate speech detection. All models were evaluated on the same Kurdish test set for fair comparison.

The most striking finding is the strong performance of pure cross-lingual transfer. Training solely on combined Persian data (Combined_Persian) and testing on Kurdish yielded excellent results with traditional ML (LinearSVC: macro-F1 0.741), significantly outperforming Kurdish baseline (0.609). This suggests substantial overlap in hate speech patterns between Persian and Kurdish. In contrast, mixing raw Persian with original Kurdish led to a drop in performance. The small proportion of Kurdish data (3%) appears to have introduced noise rather than a helpful signal. Table 5 summarises the experiments described in this section.

Translation-based augmentation produced mixed outcomes. The MT_Kurdish_Aug setting (translated Persian + original Kurdish) achieved a modest improvement for some models (e.g., ParsBERT reached 0.640), but overall gains were limited compared to Persian-only cross-lingual transfer. Translation-only data (MT_Kurdish) performed worse than the Kurdish baseline in traditional ML, indicating that current machine translation quality still introduces noticeable noise. Despite manual post-editing by a native speaker, translation-based configurations consistently underperformed pure Persian-only transfer. We attribute this to residual translationese, loss of culture-specific hate expressions, and the morphological complexity of Kur-

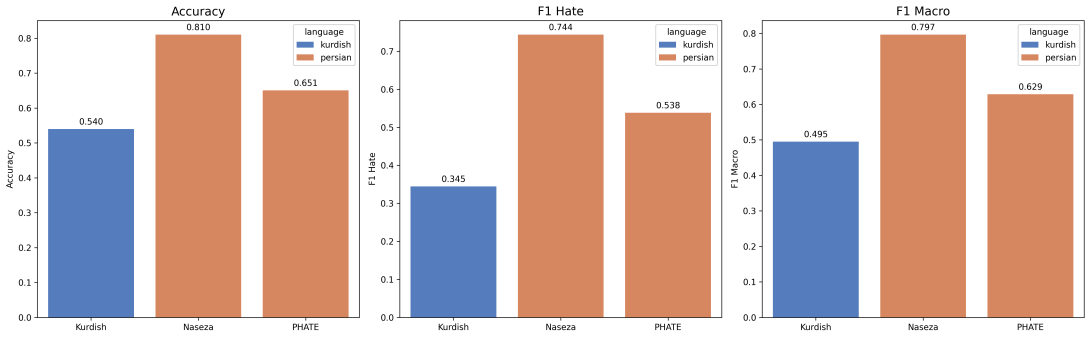

Figure 1: Rule-based approach (Paradigm A) performance across small datasets

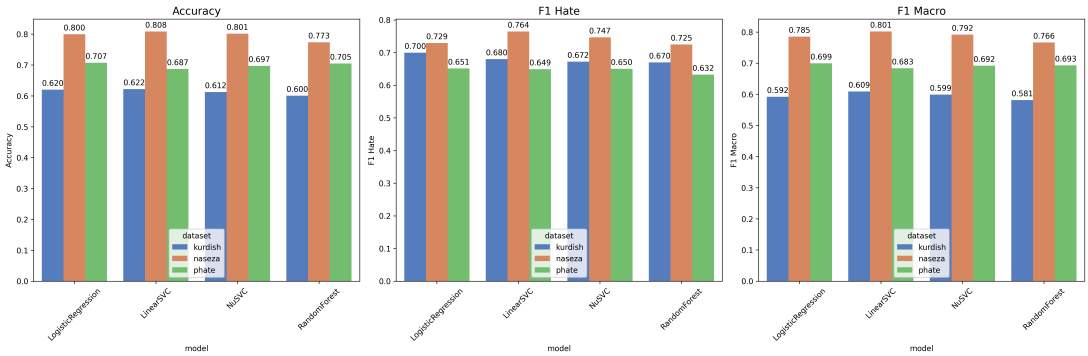

Figure 2: Traditional ML TF-IDF-based models (Paradigm B) - Phase I

Dataset	Model	Best macro-F1	Prompt
Kurdish	Qwen3-8B / Gemma-12B	0.525, 0.499	Few-shot
Naseza	Gemma-12B	0.863	Few-shot
PHATE	Gemma-12B	0.770	Few-shot

Table 4: Best macro-F1 scores for LLM paradigm (Paradigm D - Phase I)

Dataset	Training Data Composition	Best Model	Best macro-F1
combined_persian	Persian data only (Naseza + PHATE + PHICAD)	LinearSVC	**0.741**
combined_kp	Persian + original Kurdish combined	LinearSVC	0.690
mt_kurdish_aug	Original Kurdish + Persian-to-Kurdish translation	RandomForest	0.608
mt_kurdish	Persian-to-Kurdish translation only	RandomForest	0.572
combined_kp_mt	Full combination (Persian + original Kurdish + translation)	LogisticRegression	0.561

Table 5: Best macro-F1 scores for traditional ML TF-IDF-based models (Paradigm B - Phase II)

dish that current machine translation systems (even after post-editing) fail to capture fully. This finding suggests that for closely related languages, model-level transfer on original data can be more robust than imperfect data-level augmentation.

The full combination setting (Combined_KP_MT) yielded the lowest performance among Phase II configurations, further confirming that unrefined addition of translated data can be counterproductive for some model types.

In Phase II, LLM results mirrored transformer trends. Gemma (few-shot) achieved the best performance on the augmented dataset (MT_Kurdish_Aug) with a macro-F1 of 0.642. Persian-only cross-lingual transfer also performed strongly, while translation-heavy configurations

(MT_Kurdish and Combined_KP_MT) generally underperformed, signalling that raw Persian data provides a more reliable signal than current machine-translated data for Kurdish hate speech detection.

7 Conclusion and Future Work

This study systematically examined hate speech detection in Persian and Kurdish through multiple paradigms. Phase I established language-specific baselines, while Phase II demonstrated that Persian data can effectively support Kurdish detection via cross-lingual transfer. Notably, training on Persian data alone often outperformed Kurdish baselines, with ParsBERT consistently delivering the strongest results. In line with recent studies, the best performance was achieved by domain-specific deep learning models (transformers), followed by traditional ML models, both of which outperformed LLMs. Translation-based augmentation provided limited gains and sometimes introduced noise.

Future Directions. Future work should focus on developing better Persian-to-Kurdish translation models, exploring advanced cross-lingual techniques such as meta-learning and adapters, creating larger native Kurdish datasets, and incorporating multimodal signals for more robust real-world moderation systems.

Limitations. The study used a single random seed, and translation quality remained a bottleneck for data-level transfer. Domain differences across datasets and limited tokeniser coverage for Kurdish (especially in mBERT) may also affect generalisability. We did not explore re-balancing the Persian–Kurdish ratio through down-sampling of Persian or up-sampling of Kurdish; future work should investigate optimal mixing ratios. Similarly, the underperformance of LLMs may stem from tokeniser limitations on Kurdish morphology and script, domain mismatch with social-media hate speech, and prompt sensitivity, avenues we leave for future investigation.

Ethics Statement. This work focuses on hate speech detection in Persian and Kurdish, languages spoken in politically sensitive regions. We acknowledge the dual-use risk of such systems for censorship or disproportionate targeting of ethnic or political groups. All datasets used are publicly available and were collected for research purposes. We recommend that any deployment of these models incorporate human-in-the-loop moderation, regular bias audits, and alignment with local cultural and legal definitions of hate speech rather than solely Western frameworks.

Acknowledgments

This research has been partially funded by the University of Alicante, the University of Havana, the Generalitat Valenciana, and the Spanish Ministry of Science and Innovation. At the regional level, the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) and the European Regional Development Fund (ERDF) provided funding through CIDEGENT (CIDEXG/2023/13). At the national level, support was provided through the HEART-NLP project (PID2024-156263OB-C22).

References

Sina Ahmadi, Daban Q. Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045, Torino, Italia. ELRA and ICCL.

Ghadeer Al-Badani, Muneer Hazaa Alsurori, and Akram Alsubari. 2025. Transfer learning and cross-linguistic generalization in multilingual hate speech detection: Approaches and challenges. In *2025 5th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pages 1–8.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl D. Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Wazir Ali and Sampo Pyysalo. 2024. A survey of large language models for european languages. *CoRR*, abs/2408.15040.

Ahmad Amirivojdan. 2025. Naseza: A large-scale dataset for persian hate speech and offensive language detection.

K.M. Awlla, H. Veisi, and A.A. Abdullah. 2025. Sentiment analysis in low-resource contexts: Bert's impact on central kurdish. *Language Resources & Evaluation*, 35(1).

Soran Badawi. 2024. Deep learning-based cyberbullying detection in kurdish language. *Comput. J.*, 67(7):2548–2558.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech

detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

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

Hadi Davardoust, Hadi Zare, and Hossein RafieeZade. 2024. The dark side of instagram: A large dataset for identifying persian harmful comments. *SoCal NLP Symposium 2024*.

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. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17889–17897. AAAI Press.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Process. Lett.*, 53(6):3831–3847.

Endrit Fetahi, Arsim Susuri, Mentor Hamiti, Zenun Kastrati, Ercan Canhasi, and Arta Misini. 2025. Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. *Soc. Netw. Anal. Min.*, 15(1):82.

Anderson Almeida Firmino, Cláudio de Souza Baptista, and Anselmo Cardoso de Paiva. 2024. Improving hate speech detection using cross-lingual learning. *Expert Systems with Applications*, 235:121115.

Abhinav Reddy Gutha, Nidamanuri Sai Adarsh, Ananya Alekar, and Dinesh Reddy. 2023. Multilingual hate speech and offensive language detection of low resource languages. In *Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023*, CEUR Workshop Proceedings, pages 445–458. CEUR-WS.org.

Katharina Hämmerl, Jindrich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual alignment - A survey. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 10922–10943. Association for Computational Linguistics.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Aiqi Jiang and Arkaitz Zubiaga. 2026. Cross-lingual offensive language detection: A systematic review of dataset, approach and challenge. *ACM Comput. Surv.*, 58(10).

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

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

Kyuhan Lee and Sudha Ram. 2024. Deep learning for hate speech detection: A personality-based approach. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1667–1671, New York, NY, USA. Association for Computing Machinery.

Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, and Fumito Masui. 2023. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Inf. Process. Manag.*, 60(5):103454.

Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with Persian. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 89–95, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Soc. Netw. Anal. Min.*, 14(1):244.

Homa Nasiri and Morteza Analoui. 2022. Persian stance detection with transfer learning and data augmentation. In *2022 27th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–5.

Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2023. Tackling hate speech in low-resource languages with context experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, ICTD '22, New York, NY, USA. Association for Computing Machinery.

Aillkeen Bezerra de Oliveira, Cláudio de Souza Baptista, Anderson Almeida Firmino, and Anselmo Cardoso de Paiva. 2023. Using multilingual approach in cross-lingual transfer learning to improve hate speech detection. In *ICEIS (1)*, pages 374–384.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Filippo Pallucchini, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. 2026. Lost in alignment: A survey on cross-lingual alignment methods for contextualized representation. *ACM Comput. Surv.*, 58(5):116:1–116:34.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the transformer. *Soc. Netw. Anal. Min.*, 14(1):204.

Ethan Roberts. 2024. Automated hate speech detection in a low-resource environment. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 5(1).

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Ari M. Saeed, Aso N. Ismael, Danya L. Rasul, Rayan S. Majeed, and Tarik A. Rashid. 2022. Hate speech detection in social media for the kurdish language. In *Proceedings of the ICR'22 International Conference on Innovations in Computing Research, Athens, Greece, 29-31 August, 2022*, Advances in Intelligent Systems and Computing, pages 253–260. Springer.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 24(3):25:1–25:44.

Mohammad Karami Sheykhlan, Jana Shafi, Saeed Kosari, Saleh Kheiri Abdoljabbar, and Jaber Karimpour. 2023. Pars-hao: Hate and offensive language detection on persian tweets using machine learning and deep learning.

Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. 2022. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*, 101:108005.

Gemma Team. 2025a. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025b. Qwen3 technical report. *CoRR*, abs/2505.09388.

Muhammad Usman, Muhammad Ahmad, M Shahiki Tash, Irina Gelbukh, Rolando Quintero Tellez, and Grigori Sidorov. 2025. Multilingual hate speech detection in social media using translation-based approaches with large language models. *arXiv preprint arXiv:2506.08147*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2020, Online, July 9, 2020*, pages 120–130. Association for Computational Linguistics.