Culture Matters in Toxic Language Detection in Persian

Anonymous ACL submission

Abstract

Toxic language detection is crucial for creating safer online environments and limiting the spread of harmful content. While toxic language detection has been under-explored in Persian, the current work compares different methods for this task, including fine-tuning, data enrichment, zero-shot and few-shot learning, and cross-lingual transfer learning. What is especially compelling is the impact of cultural context on transfer learning for this task: We show that the language of a country with cultural similarities to Persian yields better results in transfer learning. Conversely, the improvement is lower when the language comes from a culturally distinct country.

1 Introduction

Toxic language detection focuses on identifying and mitigating harmful content in text, including but not limited to hate speech, harassment, and threats (Hoang et al., 2024). With the rapid growth of online platforms and forums, the prevalence of such toxic language has become a pressing concern. Engaging in online discussions on social media, blogs, or comment sections often exposes users to hostile or disrespectful interactions (Olteanu et al., 2018). Such toxic behaviors not only undermine the overall quality and inclusivity of online communities but are also deeply intertwined with cultural and linguistic norms. What is considered toxic or inappropriate varies significantly across cultures (Sap et al., 2021; Zhou et al., 2023b), adding complexity to the task of automatic detection.

Over the years, studies have explored various techniques for tackling the challenge of detecting toxic language across diverse languages (Abro et al., 2020; Zimmerman et al., 2018; Badjatiya et al., 2017; Gaydhani et al., 2018). Since Large Language Models (LLMs) have demonstrated outstanding performance across diverse language-related tasks in multiple languages, there is increasing interest in assessing their effectiveness in detecting toxic content. (Khondaker et al., 2023; Kumar et al., 2024; Abaskohi et al., 2024).

However, toxic language detection in Persian remains under-explored, primarily due to the lack of high-quality datasets and tailored tools. Persian (also known as Farsi) and its variants—Dari and Tajik—are spoken by over 110 million people worldwide, with significant linguistic and cultural importance[1]. Addressing the challenges of toxic language detection in Persian is crucial due to its widespread use and the complexities introduced by its non-Latin script, diverse writing styles, and regional dialects.Only a recent work by Delbari et al. (2024) showed that advanced models, such as chatGPT, struggle with detecting hate-speech in Persian, while the best performance using a fine-tuned Persian BERT model achieves only 0.61 F-Score.

To bridge this gap, our study investigates various approaches for Persian toxic language detection, including fine-tuning multiple LLMs, data enrichment, zero-shot and few-shot learning, and cross-lingual transfer learning. A key insight from our work is highlighting the critical role of cultural context in enhancing transfer learning effectiveness. Our findings indicate that models trained on languages of countries with greater cultural similarities to Persian achieve superior performance in detecting toxic content compared to those trained on large-scale English datasets, which offer only marginal improvements. This highlights the role of cultural similarities in improving model effectiveness, especially in context-dependent approaches.

Through this study, the current study addresses four research questions (RQs):

RQ1. What is the performance of existing generative LLMs on toxic language detection in Persian, using zero-shot and few-shot learning?

RQ2. Can fine-tuning enhance the performance?

[1] https://www.ethnologue.com/

RQ3. Would data enrichment via distant supervision improve Persian toxic language detection?

RQ4. Given that toxic speech classifiers are culturally insensitive (Lee et al., 2023), can cross-language transfer learning improve performance? Which languages perform best?

We explore these research questions through experiments on the PHATE dataset (Delbari et al., 2024), which includes three categories of toxic language in Persian: hate, vulgarity, and violence. For consistency, we utilize the same training, validation, and test splits as provided in PHATE. We find that toxic language identification in Persian continues to be a challenging task for most existing LLMs. However, between ParsBERT (Farahani et al., 2021) and Dorna2-Llama3 Instruct (PartAI, 2024), the two models specifically trained on Persian, Dorna2-Llama3 Instruct yield better overall results, also outperforming other multilingual models such as XLM-R and mT5. In addition, using distant supervision to obtain additional Persian training data significantly enhances the performance of ParsBERT compared to other models. We also find that transfer learning for Persian toxic detection is highly dependent on cultural context. Specifically, when the source and destination languages originate from culturally overlapping countries, the results tend to improve significantly.

2 Related Work

2.1 Toxic Language Detection

Early toxic language detection research focused on ML and DL techniques for English hate speech on social media (Asogwa et al., 2022; Davidson et al., 2017; Mullah and Zainon, 2021; Malik et al., 2024; Zimmerman et al., 2018; Zhou et al., 2020; Roy et al., 2020; Zhang et al., 2018), alongside efforts in offensive language (Bade et al., 2024; Aiyanyo et al., 2020; Cao et al., 2020; Risch et al., 2020) and cyberbullying detection (Wang et al., 2020; Pamungkas and Patti, 2019; Van Hee et al., 2015; Guo and Gauch, 2024; Cano Basave et al., 2013). Research has since expanded to languages like Indonesian (Ibrohim and Budi, 2019), Danish (Sigurbergsson and Derczynski, 2020), Arabic (Mubarak et al., 2021; Bensalem et al., 2023), Korean (Jeong et al., 2022), Chinese (Deng et al., 2022), Greek (Pitenis et al., 2020), and Indic languages (Gupta et al., 2022), including Hindi (Kapoor et al., 2019).

With LLMs, benchmarking across languages has further advanced the field (Zampieri et al., 2020; Verma et al., 2022; Caselli et al., 2021; Saleh et al., 2023; Nguyen et al., 2023; Chiu et al., 2021; Zampieri et al., 2023). Studies such as (Vargas et al., 2023), (Lu et al., 2024), and (Hoang et al., 2024) have demonstrated promising results for English. Shared tasks, like SemEval OffensEval (Zampieri et al., 2019), HASOC (Mandl et al., 2019), OSACT5 (Mubarak et al., 2022), and GermEval (Wiegand et al., 2018), have fostered collaboration and innovation in this field.

However, research on Persian toxic language detection remains rare. Existing studies (Jey et al., 2022; Sheykhlan et al., 2023; Safayani et al., 2024; Ataei et al., 2023) provide limited publicly available datasets and primarily focus on a single category of toxic language. Recently, Delbari et al. (2024) provides a hierarchical, multi-label dataset categorizing violence, hate, and vulgarity, which forms the foundation of our work. The study evaluated different models, including ParsBERT, mBERT, XML-R, and ChatGPT, with the F1-Macro of 57.8, 55, 58.3, and 43.5 respectively. Because this work uses a limited dataset, relies solely on fine-tuning BERT-base models, with GPT models restricted to zero-shot scenarios, focuses only on binary classification tasks, and lacks thorough error analysis, we aim to address these limitations by enhancing the dataset with distant supervision, experimenting with various LLMs and transfer learning techniques considering the role of cultural similarities and expanding from binary to multi-class classification to better capture real-world complexities. Additionally, we establish a robust benchmark and perform comprehensive error analysis, offering deeper insights and a more reliable evaluation framework.

2.2 Transfer Learning

Transfer learning leverages pre-trained models to improve performance on new tasks with limited data. Understudied languages can benefit significantly from this technique, as pre-trained models provide a strong foundation for adaptation and learning (Unanue et al., 2023), even though they may yield suboptimal results for tasks that rely heavily on context and culture, (Zhou et al., 2023b). Bigoulaeva et al. (2021) uses cross-lingual transfer learning for hate speech detection, leveraging English as the source and German as the target language. The approach successfully achieves strong performance on the target language without requiring annotated German data. Another study (Zhou et al., 2023a) focuses on detecting offensive language in Chinese using transfer learning with data from English and Korean. It finds that culture-specific biases hinder effective transferability.

2.3 Weak Supervision Annotation

Distant supervision is a weak supervision method that automates the creation of labeled training data by aligning unstructured text with existing annotated data. Magdy et al. (2015) demonstrates how distant supervision can assign YouTube video categories as labels to tweets linking those videos, enabling the generation of a large, automatically labeled dataset. Similarly, Go et al. (2009) applied this method for Twitter sentiment classification, achieving promising results. Additionally, studies such as (Lin et al., 2022), (Zeng et al., 2015), (Purver and Battersby, 2012), and (Mintz et al., 2009) have successfully deployed distant supervision across various NLP tasks, further showcasing its effectiveness. In this study, we introduce, for the first time in Persian, a novel distant supervision method to enhance the existing dataset.

3 Dataset

PHATE dataset, (Delbari et al., 2024), used in our study, consists of 7,056 tweets distributed across four classes: 582 labeled as violence, 1,583 as vulgar, and 1,632 as hate, with the remaining 3,259 categorized as neutral. The annotation methodology adopted in the baseline defines 'hate speech' as any instance labeled under vulgarity, violence, or hate, resulting in overlapping labels. Since our goal is distinct multi-class categorization rather than binary classification, we removed this overlapping label to concentrate on distinct toxic categories. (We evaluate all models on the same test set and adhere to the baseline train-test-validation split (50-40-10) for comparability.)

To apply distant supervision, we first needed to construct a Persian toxic lexicon. To this end, three native Persian speakers meticulously analyzed the training dataset to identify keywords frequently used in each category. This initial review yielded 164 keywords, which we refined to 127 by removing terms that could appear in neutral contexts, such as specific names, to reduce potential bias. The final selection was determined through majority voting among the annotators. At this point, nearly 40% of the keywords were associated with vulgarity.

We then followed a structured approach for each toxic class to further expand the lexicon. To enrich the "hate" category, we relied on definitions from the baseline annotation guidelines (Delbari et al., 2024) and introduced annotators to the most common hate targets. To do so, in addition to the hate targets identified by Silva et al. (2016), including racial and ethnic, religious, gender, individuals with disabilities, and other social groups, we added another hate target—politics—as the frequency of this target is reported to be high in the dataset (Delbari et al., 2024). Inspired by (Grimminger and Klinger, 2021), we also selected specific critical cultural events and asked annotators to generate keywords associated to hate speech based on those events. This approach ensured more contextually relevant hate speech categories, tailored to the sociocultural climate of the region. Annotators were asked to add relevant keywords associated with these targets, leaving categories blank where no suitable terms were identified. This process produced 216 distinct keywords, which were then narrowed down to 118 through majority voting. Next, for "violence" category, the annotators used the baseline definitions to identify relevant terms, ultimately finalizing 81 distinct keywords. Since the vulgarity class already had substantial representation, we supplemented it with 51 additional keywords at this stage.

To enhance the lexicon, we use the FastText model (Bojanowski et al., 2017) trained in Persian to identify related and synonymous terms for the 377 keywords identified earlier. Filtering out duplicates and irrelevant words, yeilded a final lexicon of 604 toxic keywords across the three categories.

Using this toxic lexicon and a Twitter archive[2], containing tweets from 2011 to 2022, we identified tweets that included the identified toxic keywords. These tweets were then labeled according to the respective categories in our lexicon. To ensure that our dataset remained distinct from the baseline dataset, which focuses on tweets from 2020 to 2023, we excluded any repeated tweets from this overlapping time frame before starting the labeling.

Ultimately, this process yielded 3291 toxic tweets across the three categories. To keep the dataset fairly balanced, we supplemented this with 3,200 neutral tweets. Tweets were considered neutral if they did not contain any of the toxic keywords from our lexicon.

[2] https://archive.org/details/twitterarchive

4 Experimental Setup

Based on the results of the recent study by Delbari et al. (2024), we selected ParsBERT (Farahani et al., 2021) as our baseline model, as it has demonstrated promising results across a variety of Persian NLP tasks. Table 4 in the Appendix lists LLMs used in our benchmarking process. All models were trained for 10 epochs on PHATE, and the final results on the baseline test dataset are from the epoch with the highest F1 score on the validation set. This methodology ensures that we capture each model's optimal performance during evaluation.

4.1 Zero/Few shot Experiments:

In our experiments, we conducted few-shot and zero-shot evaluations with Llama 3 and GEMMA 2. However, due to their poor and non-competitive performance, we excluded these results from the benchmark. We employed GPT 3.5 Turbo in both zero-shot and few-shot settings to compare performance across each class, and a binary classification setting to evaluate whether the model performs better in binary or multi-label tasks. Inspired by prior work (Abaskohi et al., 2024), We exclusively used English prompts, as they have consistently demonstrated better performance for various Persian tasks. Our prompt provides definitions for each label, based on the definitions presented in (Delbari et al., 2024), which are partially derived from Twitter's rules and policies.

4.2 Fine-tune Experiments:

We fine-tuned different LLMs on the enriched and baseline train datasets and evaluated their performance on the baseline test set, to maintain comparability. This allowed us (1) to assess the effectiveness of our distant supervision method in enriching the toxic dataset, and (2) to benchmark the performance of different state-of-the-art LLMs on the task of toxic content detection in Persian. Among our experiments on multilingual LLMs, Llama 3 consistently achieved better results compared to other models. Motivated by these findings and inspired by (Abaskohi et al., 2024), we conducted an additional experiment by translating the baseline training dataset into English using the Google Translate API. We then fine-tuned Llama 3 on the translated dataset and evaluated its performance using the baseline test set to analyze the impact of language translation on classification results. This step underscores Llama 3's adaptability and robustness across different languages.

4.3 Transfer Learning Experiments

Regarding transfer learning, we utilized three languages—Arabic, Indonesian, and English— and explored the interplay of linguistics and cultural factors in toxic speech detection. Since Llama 3 consistently achieved better results compared to other multilingual models, we selected this model for all our transfer learning experiments.

Arabic, a Semitic language, is commonly used for communication throughout the Arab world. It is written in the Arabic script and is known for its rich structure, complex grammar, and variety of regional dialects. Arabic was included in this study due to its cultural and linguistic similarities with Persian, as both languages share certain linguistic and cultural features and use similar scripts.

English, a high-resource language with extensive datasets, allows us to assess how effectively models can adapt knowledge from a linguistically and culturally unrelated yet well-documented source.

Indonesian, or Bahasa Indonesia, is the official language of Indonesia and a standardized form of Malay. As part of the Austronesian language family, it is spoken by millions across the Indonesian archipelago. Indonesian was selected for this study due to its cultural ties with Persian, enabling an exploration of how cultural similarities and linguistic differences impact transfer learning.

Regarding Arabic, we leverage the availability of large datasets for vulgar and hate speech (Mubarak et al., 2022) to examine whether the cultural and linguistic proximity between Arabic and Persian supports this approach. In one experiment, we train the Llama 3-base model on Arabic vulgar and hate datasets and evaluate its performance on the baseline test set. In another experiment, we combine the baseline Persian training dataset with the Arabic dataset, retrain the Llama 3-base model, and test it on the baseline test set. A similar approach has been applied to English, leveraging extensive datasets containing hate, vulgarity, and violence (Kennedy et al., 2020), as well as to Indonesian, utilizing a comprehensive hate dataset (Ibrohim and Budi, 2019). Finally, we conducted two additional experiments: first, by combining the Indonesian and Arabic datasets to retrain the Llama 3 base model and evaluating it on the baseline test set; and second, by integrating the baseline training dataset with the Indonesian and Arabic datasets and repeating the experiment. To ensure comparability, we

	Model	Violence			Hate			Vulgar			
		P	R	F_1	P	R	F_1	P	R	F_1	F_{macro}
Zero/Few shot	GPT 0-shot	35	75	48	39	**89**	54	61	**46**	52	51
	GPT 2-shot	40	**81**	54	55	69	61	79	37	50	55
	GPT 0-shot binary	**81**	73	**77**	**83**	64	72	**85**	30	44	64
	GPT 1-shot binary	80	70	75	77	83	**80**	74	43	54	69
	GPT 2-shot binary	78	75	76	74	86	**80**	77	42	**55**	**70**
	GPT 3-shot binary	79	71	75	76	81	78	76	36	49	67
Fine tuning	ParsBert (Baseline)	68	42	52	**63**	59	60	55	**68**	**60**	57
	Dorna2-Llama Inst.	61	**74**	**67**	56	73	60	50	52	55	**61**
	XLM-R-base	63	50	56	58	67	**62**	55	63	59	59
	Llama 3 - Base	68	57	62	53	**76**	**62**	51	65	57	60
	Llama 3 translated	48	57	52	49	67	57	36	34	35	48
	Llama 3 Instruct	**74**	55	63	59	55	57	58	57	57	59
	GEMMA 2	57	35	43	51	69	59	40	54	46	49
	mT-5	38	41	39	56	49	52	**59**	26	36	42
Distant supervision	ParsBert	**62**	58	60	**71**	**81**	**75**	**78**	**67**	**72**	**69**
	XLM-R	54	69	**61**	**71**	74	72	76	63	69	67
	Llama 3	36	**70**	47	70	57	63	56	51	53	54
	Gemma 2	37	65	47	64	54	58	44	50	47	51
	mT-5	34	61	44	45	74	56	52	62	57	52
Transfer learning	Llama 3 - En	78	69	73	55	60	57	74	81	77	69
	Llama 3 - En+Fa	**79**	**70**	**74**	56	61	59	81	78	80	71
	Llama 3 – Ar	-	-	-	75	89	81	81	**84**	82	82
	Llama 3 – Ar+Fa	-	-	-	86	88	87	**83**	**84**	**84**	**86**
	Llama 3 – Id	-	-	-	89	84	86	-	-	-	-
	Llama 3 – Id+Fa	-	-	-	92	80	86	-	-	-	-
	Llama 3 - Ar+In	-	-	-	**94**	**92**	**93**	-	-	-	-
	Llama 3 - Ar+In+Fr	-	-	-	92	91	91	-	-	-	-

Table 1: Toxic detection across approaches. Best in each group bolded, Overall best underlined.

used datasets of equal size for all languages while maintaining a balanced label distribution across all classes. We achieved this by randomly selecting an equal amount of data from each dataset. Due to the absence of large datasets for vulgar or violent language, our Indonesian experiments focused solely on hate detection. Similarly, the lack of Arabic and Indonesian datasets for violence restricted our transfer experiments to English.

5 Results

This section is divided based on the results obtained using different methods as Zero-Shot/Few-Shot, Fine-Tuning, Distance Supervision, and Transfer Learning approach.Table 1 presents a comprehensive comparison of model performance.

5.1 GPT 3.5 Turbo Few-Shot and Zero-Shot

For multi-class classification, GPT 3.5 Turbo - 0 Shot achieved moderate scores across categories, while GPT 3.5 Turbo - 2 Shot improved these metrics, notably for Hate and Violence. However, increasing the number of shots beyond two did not yield significant improvements in performance. To optimize resource utilization, we limited our experiments to 2-shot settings for multi-class classification and shifted our focus to binary classification for further evaluation. In binary classification, models demonstrated significantly higher performance overall. GPT 3.5 Turbo - 0 Shot achieved top scores in categories such as "Violence" and "Hate".

5.2 Fine Tuning

The fine-tuning results revealed distinct trends among the four LLMs groups.

BERT Models: ParsBERT, the BERT-base model, served as the baseline (Delbari et al., 2024) achieved moderate F1 scores for all categories. When fine-tuned with an enriched dataset, ParsBERT with Distant Supervision showed significant improvements on the baseline test set particularly for "Hate" (F1 = 75) and "Vulgar" (F1 = 72). Additionally, the performance of the XLM-R-base model, fine-tuned with the enriched dataset, improved significantly across all categories.

Llama Models: The Llama models displayed varied performance depending on the dataset and specefic models. Llama 3 – Base, trained on the baseline dataset, achieved F1 scores of 62, 62, and 57 for "Violence," "Hate," and "Vulgar," respectively. However, its enriched counterpart, Llama 3 with Distant Supervision, showed mixed results: while the F1 score for "Hate" improved, the score for "Violence" dropped significantly, highlighting challenges in effectively utilizing enriched datasets. A similar drop occurred for "Vulgar." Compared to other models, Llama 3 – Translated, fine-tuned on English-translated baseline dataset, underperformed, suggesting that translation into English may have removed critical linguistic features necessary for effective classification. Llama 3 – Instruct trained on the baseline training dataset achieved consistent F1 scores of 63, 57, and 57 across the three categories. Building on these findings, we extended our experiments by incorporating the recently released Dorna2-Llama3 Instruct, which outperformed both Llama 3 – Instruct and Llama 3 – Base, achieving higher F1 scores for the 'Violence' and 'Hate' classes. Notably, among all fine-tuned models in our experiments, this model achieved the highest results for detecting 'Violence'.

GEMMA Models: The GEMMA 2 models, underperformed compared to Bert - base and Llama - base models. Enriching the dataset offered marginal improvements for "Vulgar" but for "Violence" increased 4% and "Hate" dropped by 1%. These results highlight the limitations of GEMMA 2 in task-specific Persian contexts.

mT-5 Model: mT-5 exhibited the weakest performance among all fine-tuned models. While mT-5 with Distant Supervision showed slight improvements, it struggled to achieve competitive results.

5.3 Transfer Learning

We observed that fine-tuning on English data alone (Llama 3 – Eng) yielded moderate results: While the model performed well in "Violence" and "Vulgar," its performance in "Hate" was weaker. Including Persian in the training process alongside English (Llama 3 – Eng + Fa) improved the F1 scores across all categories.

Fine-tuning on Arabic data alone (Llama 3 – Ar) yielded strong F1 scores of 81 for both "Hate" and "Vulgar." Adding Persian data (Llama 3 – Ar + Fa) further enhanced performance, with F1 scores of 87 for "Hate" and 84 for "Vulgar."

Fine-tuning on Indonesian alone (Llama 3 – Id) resulted in an F1 score of 86 for 'Hate.' However, incorporating Persian data into the Indonesian training set (Llama 3 – Id + Fa) further improved precision while maintaining a consistent F1 score.

Integrating both Arabic and Indonesian datasets (Llama 3 - Id + Ar) achieved the highest F1 score of 93 across all experiments. However, adding Persian (Llama 3 - Id + Ar + Fa) resulted in a slight decrease, bringing the F1 score down to 91.

6 Analysis and Discussion

6.1 RQ1: Generative LLMs Performance

Our first RQ concerned the performance of existing generative LLMs, using zero-shot and few-shor learning: We observed that in zero- and few-shot settings, GPT-3.5 Turbo performs significantly better in binary classification tasks than in multi-label classification. In zero-shot multi-label classification, the model frequently mislabeled instances, often confusing categories such as 'hate' and 'violence.' Additionally, some instances of 'hate' are incorrectly classified as 'neutral,' particularly when lacking sufficient contextual cues.

Analysis of few shot multi-label classification reveals misclassifications that even though they contain elements of vulgarity or violence such as keywords like "down with" and "dead to", or discussions about public figures and specific locations do not meet the criteria for hate speech. Moreover, as in zero-shot multi-label classification, some instances of 'hate' are misclassified as 'neutral,' especially those related to specific events. Table 2 shows some GPT 3.5 Turbo misclassified samples.

Given GPT 3.5 Turbo's stronger performance in binary settings, we conducted three few-shot experiments with 1-shot, 2-shot, and 3-shot settings, with noticeably better performance. After analyzing the errors in the binary setting, we found that GPT-3.5 Turbo similar to multi-classification experiments relies heavily on contextual clues in the text to distinguish between these labels. However, the predictions can skew incorrectly when the context is ambiguous or conceptually overlapping. For example, while the model successfully detects hate with common targets (e.g., religion, politics), it struggles to detect hate for targets related to specific events. Table 6 in the Appendix presents some of these misclassifications. Interestingly, the model's performance either remained steady or dropped as the number of shots increased. Analysis reveals that instances relying on context struggle to predict correctly, even in a 3-shot setting. This finding aligns with prior work that conducted exhaustive experiments on GPT models across various Persian tasks (Abaskohi et al., 2024). For N-shot binary and multi-class classification, we tested various instances at each level and selected the average-performing outputs for reliability.

6.2 RQ2: Fine-Tuning Effect

Our Second RQ concerned fine tuning: Regarding models specifically trained on Persian, in comparison to others, ParsBERT still lagged in detecting toxic language. In contrast, the recent Dorna2-Llama3.1-Instruct achieved better overall results.

Regarding multilingual LLMs, Llama 3 performs better than GEMMA 2, with mT5 being the worst among them. We also used Llama-Instruct with a definition of the classification task, but did not observe significant differences in performance. Using the translated dataset, we observed that all metrics dropped notably: likely due to the problematic translations, As most entries were informal and context-dependent, they were difficult for Google Translate to process correctly.

6.3 RQ3: Data Enrichment via Distant Supervision

Our third RQ concerned the effect of data enrichment via distant supervision: Our results demonstrates that distant supervision improves mT5 and significantly enhances BERT base models. However, it performs poorly on Llama 3 and GEMMA 2. The metrics reveal that the results on Llama 3 are 50% worse than those on GEMMA 2, suggesting that Llama-3 is less tolerant to noise when trained on Persian. Additionally, our proposed dataset introduces a drop in precision for detecting violence across all models.

Tweet (original + English translation)
رشتو در کشورهایی نظیر میهن ما ایران که مردم یا بحقوق مدنی خودشان واقف نیستند یا برای احقاق آن پای ایستادگی ندارند بگذریم که از گروه بیشماری که با ساندیس و ساندویچی به آن حقوق چشم میپوشند و فریادهای مرگ بر ..یا الله اکبر خایمه ای رهخر سرمیدهند، صحبت جمهوری خواهی پس از براندازی👇 The thread in countries like ours, Iran, where people are either unaware of their civil rights or do not stand firm to claim them — let alone the countless groups who, in exchange for a juice and a sandwich, turn a blind eye to those rights and chant slogans like **'Death to**...' or 'Allah Akbar Khamenei Rahbar' — speaks of republicanism after the overthrow 👇.
به دشمنان #رائفی_پور و #حسن_عباسی میگیم بیاید مناظره میگن احمق نیستیم که وقتمونو با یه مشت بی سواد تلف کنیم اصلا به حرفای اینا اهمیت نمیدیم!!😏بعد از اون طرف دونه دونه سخنرانی هاشونو آنالیز میکنن و از تو سخنرانیای صدسال پیششون فلان مثلا سوتی رو درمیارن یا فلان حرف واسه شکایت😏 ##**Rafi_poor** and #**Hassan_Abbasi** are in the midst of an idiotic debate when they come to me with a black face that literally means nothing!! 😏 after that party Don't say anything about someone, for example, say something about someone, for example, say something like a letter with a complaint😏
نیمار اومده به بازیکن ژاپنی مارسی توهین نژادی بکنه به جای اینکه بگه ژاپنی گوه گفته چینی گوه قشنگ ریده تو کل آسیا. به غیر از اون به بازیکن مارسی هم همجنسگرا خطاب کرده که احتمالا محرومیت سنگینی رو در انتظار داره. متاسفانه پول علاوه بر شعور سواد جغرافیایی هم نمی توان داشت Neymar insulted the **Japanese Marseille** player with a racial slur. Instead of saying 'Japanese shit,' he said **'Chinese shit**,' completely screwing over all of Asia. Besides that, he also called the Marseille player gay, which will likely result in a heavy suspension. Unfortunately, money can't buy intelligence or geographical knowledge.
رژیم صهیونیستی در واقع یک رژیمی است که پایه‌های آن بشدّت سست است، رژیم صهیونیستی محکوم به زوال است #freepalestine #مهد_مقاومت **The Zionist regime** is a regime with extremely fragile foundations; the Zionist regime is doomed to collapse.#freepalestine #Stronghold_of_Resistance

Table 2: Samples of false negative classifications by the GPT for the Hate class.

Tweet (original + English translation)	Ar	Ar+Fa	In	In+Fa	Ar+In	All
اسرائیلی ها اینقدر زیاده خواه و بی منطق اند که محمود عباس تهدید کرده ؛"چنانچه اوضاع تغییر نکند علیه رژیم صهیونیستی اقدام خواهیم کرد" Israelis are so greedy and irrational that Abbas has threatened, "If the situation doesn't change, we'll take action against the Zionist regime."	1	1	0	1	1	1
رندی به محضر فقیهی رسیدو حرکات رقص را جداجدا انجام می داد و می پرسید آیا حرام است؟ فقیه میگفت نه. پس رند شروع به رقصیدن کرد.فقیه گفت تجزیه اش خوب بود ولی مرده شور ترکیبش رو بردن **حالا حکایت این عدالت خواراست بعضیشون عیییی بچه های خوبین ولی مرده شور ترکیبشون رو برده من برم به کارای خودم برسم خدافظ #عدالتخواران#انتخابات_مجلس** A trickster went to a cleric and performed dance moves separately, asking if they were forbidden. The cleric said no. Then the trickster started dancing, and the cleric said, 'Breaking it down was fine, but damn the combination!' **This is exactly the case with these so-called justice-seekers—some of them are actually good kids, but damn their combination! Anyway, I'll get back to my own business. Goodbye. #JusticeSeekers #ParliamentElection**	0	1	1	0	1	1
علیرضا دبیر: صحبت راجب سیاست شرعاً مشکل داره، از بروبچ کشتی میخوام گوشیشون رو کنار بزارن و رو تمریناتشون تمرکز کنن. بعد از انقلاب اولین نفری که میدم سگم بهش تجاوز کنه تویی بی وجود# مهسا_امینی Alireza Dabir: "Talking about politics is religiously problematic. I ask the wrestling guys to put their phones aside and focus on their training." After the revolution, the first person I'll have my dog violate is you, you worthless being. #Mahsa_Amini	0	1	1	1	1	0
ازماست که برماست تا به این دین و باورهای بیابان گرد ملخ خوار باور داریم **همین آش وهمین** کاسه It is up to us to believe in this religion and beliefs of the desert, the locust-eating locusts, **the same soup and the same bowl.**	0	1	0	1	0	1
از دی به سینا ازسینا به نجمیه چه غلطی دارن میکنن ازنجمیه هم می‌خوان ببرن امیرعلم حتما بیشرفا From Dey to Sina, from Sina to Najmieh, what the hell are they doing? Now they want to move from Najmieh Amir Alam. They must be absolute scoundrels	0	0	0	0	0	0

Table 3: Transfer Learning model predictions for Hate Farsi tweets across multiple languages.

As highlighted by (Magdy et al., 2015), distant supervision, despite its inherent noise, can substantially enhance model performance by providing additional contextual data during training. This observation aligns with our findings, where the BERT-base models demonstrated improved performance with distant supervision.

However, as Table 1 shows, for ParsBERT and XLM-R, the precision for the "violence" category dropped by an average of 7%. A detailed analysis of misclassified labels revealed that 68% of "neutral" labels were erroneously classified as "violence." This misclassification can primarily stemmed from overlapping keywords and contextual ambiguities triggered by our toxic lexicon. For example, in the enriched dataset, the word بمر (kill) often appears in both "neutral" and "violent" contexts. While in Persian it is typically used humorously or exaggeratedly in neutral conversations, the models frequently misclassified it as "violent". Similarly, terms like موشک زدند (barrage rocket) and collocation with منفجر(explode), neutral in certain contexts, were incorrectly labeled as violence. Table 5 displays some of the false positive instances resulting from the model. Since most of these tweets were correctly labeled as neutral during the baseline training of the BERT-base models, this suggests that our distant supervision method introduced noise, complicating the differentiation between categories in this context.

In addition, we observed that, although the instances for the "vulgar" category increased by approximately 40% through distant supervision, the recall remained almost unchanged for both ParsBERT and XLM-R. This stability in recall suggests that the additional data introduced by distant supervision might not have been sufficiently diverse or contextually rich to enhance the models' performance. Moreover, the models still struggle with implicit profane speech. Table 5 in the Appendix presents instances that were not detected as 'vulgar' during training on both datasets, even though they explicitly contain words from our toxic lexicon. In contrast, our dataset significantly improves the recall for "hate". We observed that this is especially true for hate directed towards politics, where the model trained on the baseline dataset struggled to identify instances. However, after training on the enriched dataset, it successfully detected these instances, suggesting that our approach for identifying hate keywords in the toxic lexicon works well for hate detection.

6.4 RQ4: Cross-Lingual Transfer Learning

Our fourth RQ concerned the effect of culture in transfer learning: Our findings indicate that while Persian can effectively benefit from the Arabic and Indonesian datasets, its performance gains from the English dataset are less pronounced. Closer analysis of the results suggests two potential reasons for this disparity. First, the general culture of hate in Persian, Arabic, and Indonesian appears to be more similar, particularly in targets related to religion, politics, and common controversial events that provokes hate. In contrast, the English hate dataset predominantly focuses on contexts diverging significantly from the Persian hate dataset (e.g. sexual orientation and ethnic groups). Second, both Persian and Arabic are morphologically rich languages. This shared characteristic can allow Persian to exploit the morphological richness of Arabic during transfer learning, leveraging the capacity of LLMs to process such linguistic features effectively. The pattern observed with the hate class was mirrored in the vulgar class, where Persian again benefited more from Arabic than from English. However, to assess whether the effectiveness is more cultural or linguistic, we experimented on Indonesian, which has completely distinct linguistic features from Persian. As the results show, despite its linguistic divergence, training solely on the Indonesian dataset produced even better results than Arabic. Interestingly, our experiments demonstrated that English can still provide relevant contextual information about violence applicable to Persian.

Integrating datasets from three language pairs (Arabic-Persian, English-Persian, and Indonesian-Persian) showed improved performance metrics in the first two settings, except for a slight decline in recall for the "vulgar" class in the English-Persian combination (3%) and the "hate" class in the Arabic-Persian combination (1%). These minor drops can likely be attributed to the imbalance in data samples between the two datasets (e.g. PHATE and Indonesian). Upon further examination, we observed that, the transfer learning experiments reveal some differences in how Arabic and Indonesian datasets contribute to Persian toxic language detection. Specifically, transfer learning from Arabic data helped detect hate speech related to religious and political topics, particularly sociopolitical hate prevalent in the Middle East. This indicates that Arabic dataset provides relevant contextual cues for religious and politic discourse. On the other hand, transfer learning from Indonesian data helped detect hate speech directed toward individuals rather than groups (e.g. profession). In addition, our analysis highlights that models trained on Indonesian data exhibit significantly better performance in handling long texts containing a mix of neutral and hateful sentiments. This can be one reason Indonesian outperforms Arabic in detecting Persian hate instances. Close analysis of the Indonesian dataset, showed that it lacks sufficient political hate speech instances, which explains the model's struggle to generalize to such cases in the Persian context. Furthermore, both transfer learning approaches reveal challenges in detecting instances containing idiomatic expressions and culturally dependent references that require specific background knowledge. We observed that integrating Persian data into the training process helps mitigate these challenges for Arabic and Indonesian datasets, with a more pronounced improvement in the Arabic model. We aimed to explore whether incorporating Indonesian and Arabic datasets could improve Persian hate speech detection and whether these two languages complement each other in identifying Persian hate speech. Upon examination, we confirmed our hypothesis: the integration of these two languages effectively complemented each other, improving detection capabilities. However, when we integrated all available training data—Indonesian, Arabic, and Persian—and trained a model using this combined dataset, we observed a slight drop across all metrics, although the results remained strong. A closer error analysis failed to reveal a clear pattern explaining this decline. Further investigation is needed to determine why incorporating Persian did not lead to additional improvements. Table 3 provides sample predictions that support our findings. More examples present in Table 8 in the Appendix.

7 Conclusion

This paper presented a comprehensive evaluation of various fine-tuning, zero-shot/few-shot, and transfer learning methodologies to assess the performance of LLMs in detecting toxic content in Persian—a low-resource language. Given the limited availability of data for Persian, we explored distant supervision to enrich existing Persian datasets and transfer learning to evaluate Persian's ability to leverage resources from other languages.

Our analyses demonstrate that distant supervision significantly enhances the performance of BERT-based models, particularly ParsBERT. We also show that transfer learning is more effective when the language belongs to a country with cultural similarities to Persian, whereas improvements are less significant for languages from culturally distinct countries.

Limitations

One limitation of our study is that the toxic lexicon introduced for distant supervision cannot comprehensively capture all forms of toxic speech. Additionally, some keywords in the lexicon are heavily event-specific and may lose relevance over time as those events fade from public memory. This limitation suggests that the lexicon may not effectively identify toxic language associated with future events that provoke hate, violence, or vulgarity.

Furthermore, other forms of toxic speech, excluded due to dataset constraints, present opportunities for future research to improve toxic speech detection frameworks.

While our study focuses on only three languages, limiting broader conclusions about cross-lingual transfer learning, our selection was guided by cultural relevance to Persian. Arabic and Indonesian were chosen for their linguistic and cultural ties, while English served as a high-resource control language. Further studies should explore additional languages to enhance cross-lingual generalizability.

Ethics Statement

This study adheres to ethical principles by prioritizing the fair and responsible use of technology to detect toxic content. The methods employed are designed to minimize bias, ensure privacy, and avoid unintended harm. We emphasize the importance of transparency, accountability, and the careful consideration of societal impacts in the deployment of toxic detection systems. All data used in this research were collected and processed in compliance with relevant ethical guidelines and data protection regulations.

References

2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203, Torino, Italia. ELRA and ICCL.

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Imatitikua D Aiyanyo, Hamman Samuel, and Heuiseok Lim. 2020. A systematic review of defensive and offensive cybersecurity with machine learning. *Applied Sciences*, 10(17):5811.

Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification usingSVM and Naive Bayes. *arXiv preprint arXiv:2204.07057*.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Amin Pourdabiri, Behrouz Minaei-Bidgoli, and Mohammad Taher Pilehvar. 2023. Pars-off: A benchmark for offensive language detection on Farsi social media. *IEEE Transactions on Affective Computing*, 14(4):2787–2795.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244, St. Julian's, Malta. Association for Computational Linguistics.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Imene Bensalem, Meryem Mout, and Paolo Rosso. 2023. Offensive language detection in Arabizi. In *Proceedings of ArabicNLP 2023*, pages 423–434.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 109–117, Nagoya, Japan. Asian Federation of Natural Language Processing.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, pages 11–20.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with GPT-3. *arXiv preprint arXiv:2103.12407*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17889–17897.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The LLAMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Xiaoyu Guo and Susan Gauch. 2024. Using sarcasm to improve cyberbullying detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 52–59, Torino, Italia. ELRA and ICCL.

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

Nhat Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Anh Tuan Luu. 2024. ToXCL: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472, Mexico City, Mexico. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pegah Shams Jey, Arash Hemmati, Ramin Toosi, and Mohammad Ali Akhaee. 2022. Hate sentiment recognition system for Persian language. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 517–522.

Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9951–9952.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings* of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 220–247, Singapore. Association for Computational Linguistics.

Ankit Kumar, Richa Sharma, and Punam Bedi. 2024. Towards optimal NLP solutions: Analyzing GPT and LLaMA-2 models across model scale, dataset size, and task diversity. *Engineering, Technology & Applied Science Research*, 14(3):14219–14224.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. 2022. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863.

Junyu Lu, Bo Xu, Xiaokun Zhang, Kaiyuan Liu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Take its essence, discard its dross! debiasing for toxic language detection via counterfactual causal effect. *arXiv preprint arXiv:2406.00983*.

Walid Magdy, Hassan Sajjad, Tarek El-Ganainy, and Fabrizio Sebastiani. 2015. Bridging social media via distant supervision. *Social Network Analysis and Mining*, 5:1–12.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2024. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, pages 1–16.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 9:88364–88376.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. 2023. Fine-tuning LLAMA 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

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, Florence, Italy. Association for Computational Linguistics.

PartAI. 2024. Dorna2-llama3.1-8b-instruct. Accessed: 2025-02-01.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. 2020. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204951–204962.

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*.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

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. *Authorea Preprints*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 687–690.

Inigo Jauregi Unanue, Gholamreza Haffari, and Massimo Piccardi. 2023. T3l: Translate-and-test transfer learning for cross-lingual text classification. *Transactions of the Association for Computational Linguistics*, 11:1147–1161.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196.

Kanishk Verma, Tijana Milosevic, Keith Cortis, and Brian Davis. 2022. Benchmarking language models for cyberbullying identification and classification from social-media texts. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 26–31, Marseille, France. European Language Resources Association.

Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. 2020. Detect all abuse! toward universal abusive language detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. Offenseval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*, 29(6):1416–1435.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023b. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics:*

EMNLP 2023, pages 12684–12702, Singapore. Association for Computational Linguistics.

Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, and Nick Savage. 2020. Deep learning based fusion approach for hate speech detection. *IEEE Access*, 8:128923–128929.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

A Appendix

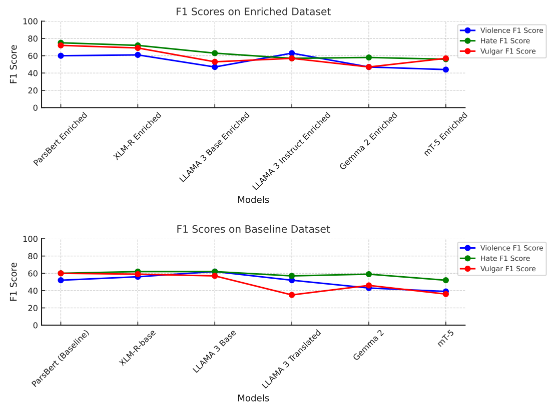


Figure 1: The fine-tuned models' performance before and after dataset enrichment.

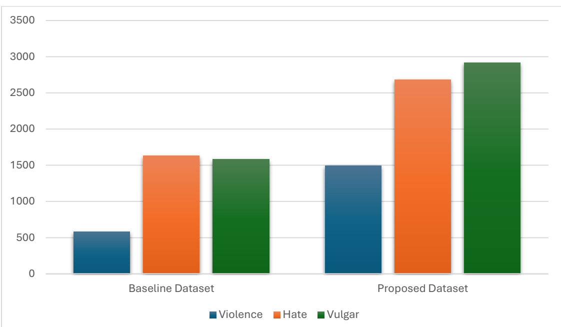


Figure 2: Label Distribution Before and After the Enrichment

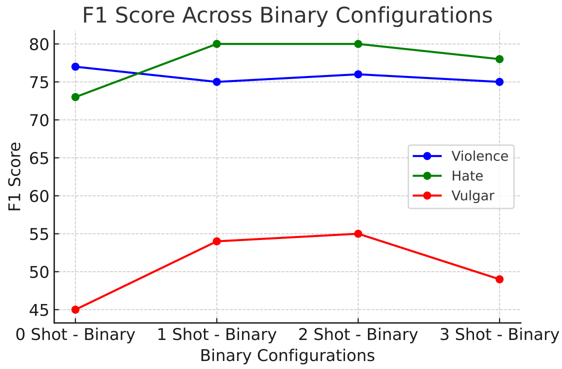


Figure 3: GPT F1 Score Results for Three Classes Across N-Shot Configurations

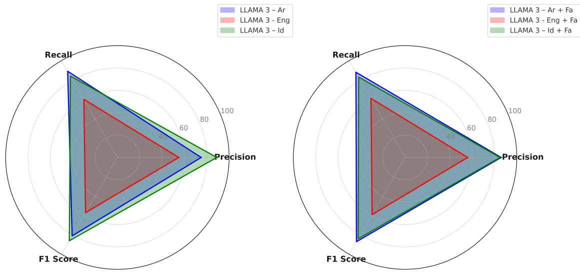
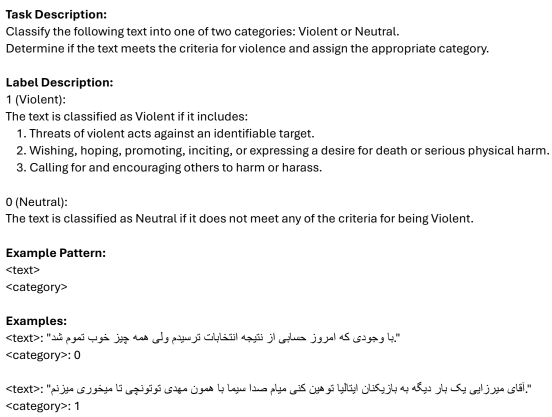


Figure 4: Performance of Transfer-Learning methods on the Test Set for Hate Speech Detection.


Task Description:
Classify the following text into one of two categories: Violent or Neutral.
Determine if the text meets the criteria for violence and assign the appropriate category.

Label Description:
1 (Violent):
The text is classified as Violent if it includes:
1. Threats of violent acts against an identifiable target.
2. Wishing, hoping, promoting, inciting, or expressing a desire for death or serious physical harm.
3. Calling for and encouraging others to harm or harass.

0 (Neutral):
The text is classified as Neutral if it does not meet any of the criteria for being Violent.

Example Pattern:
<text>
<category>

Examples:
<text>: "با وجودی که امروز حسابی از نتیجه انتخابات ترسیدم ولی همه چیز خوب تموم شد."
<category>: 0

<text>: "آقای میرزایی یک بار دیگه به بازیکنان ایتالیا توهین کنی میام صدا سیما با همون مهدی توتونچی تا میخوری میزنم."
<category>: 1

Figure 5: The Prompt Used for the GPT Experiment

Model	#Params	Reference
ParsBERT	162M	(Farahani et al., 2021)
XLM-RoBERTa-Base	125M	(Conneau, 2019)
mT5-Base	120M	(Xue et al., 2021)
Llama 3-Base	8B	(Dubey et al., 2024)
Llama 3 Instruct	8B	(Dubey et al., 2024)
Dorna2-Llama3.1-Instruct	8B	(PartAI, 2024)
GEMMA 2	9B	(tea, 2024)
GPT 3.5 Turbo	175B	(Brown, 2020)

Table 4: LLMs used in our Study.

Tweet	Actual Label	Predicted Label
زیر بارون باهم قدم بزنیم تو چترتو واسه من نگه داری که من خیس نشم ولی خودت زیر بارون خیس بشی بعد تو سرما بخوری کرونا بگیری بمیری که وقتی میگم بیا بریم خونه نگی نه بریم قدم بزنیم ((((: Let's walk together in the rain, and you hold the umbrella over me so I don’t get wet, but you get soaked in the rain. Then you catch a cold, get COVID, and die, just so the next time I say, "Let’s go home," you don’t say, "No, let’s keep walking." :)))))	Neutral	Violence
ای بمیری چقدر شکر بهش زدی Ugh, die already! How much sugar did you add to it	Neutral	Violence
وقتی کابلهای برق نطنز اتصالی کنه خو مشخصه که مرکز موشک سازی اسرائیل منفجر میشه 😉 When the power cables in Natanz short-circuit, of course, the missile manufacturing centre in Israel is going to explode 😉 .	Neutral	Violence
عمل زیبایی نه مایه شرمه نه افتخاره. (از مجموعه گه یکدیگر را نخوریم) Cosmetic surgery is neither a source of shame nor pride. (From the "Let's Not Eat Each Other" collection)	Vulgar	Hate
همون سالی که یارو حادثه رو با سریال واکنیگ دد و زامبی ها مقایسه کرد باید به عقلش شک میکردید 🤦🏾‍♂️ The year that guy compared the incident to *Walking Dead* and zombies was the moment you should’ve questioned his sanity. 🤦🏾‍♂️	Vulgar	Neutral

Table 5: Samples (with translation) of misclassification instances after training ParsBERT on enriched dataset

Tweet	Actual Label	Predicted Label				
		0-shot multi	0-shot binary	1-shot binary	2-shot binary	3-shot binary
گفتگو؟؟؟ سه ساله هرروز داریم میپرسیم #موشک_دوم رو چرا زدید Conversation??? For three years, we've been asking every day why you fired the second missile.	Hate	Violence	Neutral	Neutral	Neutral	Neutral
دختری جوان برای عمل جراحی زیبایی به کلینیکی مراجعه میکند و زیر تیغ سکته میکند؛ جسد او را به خارج بردند و آن را آتش زدند. نمیخواهید کل هیکل سازمان نظام پزشکی را از بالا تا پایین افتابه بگیرید؟ A young girl visits a clinic for cosmetic surgery and suffers a stroke under the knife; her body is taken abroad and set on fire. Don’t you want to take the whole Medical System Organization from top to bottom and throw it in the trash?	Hate	Vulgar	Hate	Hate	Neutral	Neutral
خدا رو شاکرم که علیرغم پذیرش در آزمون قضاوت و گزینش‌های مربوطه به شغل شریف قضاوت نائل نیامدم تا مجبور نباشم زمانی که پدر دو کودک ۸۰ روز در بازداشت انفرادی به سر میبرن حکم به بازداشت مادر آن‌ها نیز بدهم! I thank God that despite being accepted in the judicial exam and the related selections, I did not attain the honourable position of a judge, so I wouldn't have to give a verdict to detain the mother of two children while their father spends 80 days in solitary confinement!	Hate	Neutral	Neutral	Neutral	Neutral	Neutral

Table 6: Samples of Hate Misclassifications by the GPT Binary/Multi Classification Experiment

Language	Dataset Size			Dataset Size Used		
	Total	Hate	Not Hate	Total	Hate	Not Hate
English	39,565	10,892	28673	8050	4025	4025
Indonesian	13,169	5561	7608	8050	4025	4025
Arabic	12,698	4025	8673	8050	4025	4025

Table 7: Dataset Distribution and Subset Selection for Hate Classification in Transfer Learning

tweet	Ar	Ar + Fa	Ind	Ind + Fa	Type
بمب که بسازیم نه احتیاج به انتخابات دار_نه هیچ فشاری از طرف داخل و خارج موشک هم دارم.سپاه هم دارم. گروههای نیابتی هم دارم.اسراییل هم که آنجاست باج میگیرم و حکومت میکنم.مردم هم غلط کرده اند که به مُژتبآ روی خوش نشان ندهند..جمهوری کُرهِ شمالی اسلامی If we build a bomb, we won't need elections, nor will we face any internal or external pressure. I have missiles, I have the IRGC, I have proxy groups, and Israel is right there—I can extort and rule. And the people have no right to oppose Mojtaba. An Islamic North Korea	0	0	0	0	-Implicit hate - Sarcastic
آغاز #دهه_زجر ، آغاز اعدام بلند پایه‌ترین مقامات کشوری و لشکری میهن پرست، آغاز اعدام مردم بیگناه و آزادیخواه، آغاز سالها فقر بدبختی گرانی تورم فحشا ندانم کاری برای ایرانیان و آغازِ پایان امنیت #خاورمیانه را به تمام بی دغدغه‌ها و #جهاد_فاکتوری تبریک و به #مردم_ایران تسلیت می‌گویم. The beginning of the #Decade_of_Agony marks the start of the execution of the highest-ranking patriotic civil and military officials, the execution of innocent and freedom-loving people, the beginning of years of poverty, misery, inflation, prostitution, and incompetence for Iranians, and the start of the end of security in the #MiddleEast. Congratulations to the indifferent ones and #Jihad_Factory, and my condolences to the #People_of_Iran	1	1	0	1	Politics
اسرائیلی ها اینقدر زیاده خواه و بی منطق اند که محمود عباس تهدید کرده ؛«چنانچه اوضاع تغییر نکند علیه رژیم صهیونیستی اقدام خواهیم کرد» The Israelis are so greedy and irrational that Mahmoud Abbas has threatened, “If the situation does not change, we will take action against the Zionist regime.”	1	1	0	1	Politics
تف تو مملکتی که دیه الناز رکابی از مهدی ترابی کمتره 🫢 😟 !... Shame on a country where Elnaz Rekabi's blood money is worth less than Mehdi Torabi's...! 😟 🫢	1	1	0	1	Politics
رندی به محضر فقیهی رسیدو حرکات رقص را جداجدا انجام می داد و می پرسید آیا حرام است؟ فقیه میگفت نه. پس رند شروع به رقصیدن کرد.فقیه گفت تجزیه اش خوب بود ولی مرده شور ترکیبش رو ببرن **حالا حکایت این عدالت خواراست بعضیشون عیییی بچه های خوبین ولی مرده شور ترکیبشون رو ببره من برم به کارای خودم برسم خدافظ #عدالتخواران#انتخابات_مجلس** A trickster went to a cleric and performed dance moves separately, asking if they were forbidden. The cleric said no. Then the trickster started dancing, and the cleric said, 'Breaking it down was fine, but damn the combination!' **This is exactly the case with these so-called justice-seekers—some of them are actually good kids, but damn their combination! Anyway, I’ll get back to my own business. Goodbye. #JusticeSeekers #ParliamentElections**	0	0	1	1	Mix of Neutral and Hate (bold)
رفتم دماغمو عمل کنم دکتره یه نگاهی کرد گفت داداش شما صبر کن یکم دیگه علم پیشرفت کنه کلا سرتو عوض کن 😂 حالا شده حکایت ما مملکتی که رقابت خوشگلترین مسئولش بین جهانگیری و احمد خاتمی و احمدی نژاده باید رید توش البته از نظر عملکرد شان هم باید رید توش #مهسا_امینی #ترور_بیولوژیکی I went to get my nose done, and the doctor took one look at me and said, 'Bro, just wait a little longer until science advances enough to replace your whole head. 😂' **That’s exactly our situation—when the competition for the most handsome official in the country is between Jahangiri, Ahmad Khatami, and Ahmadinejad, you know it's doomed. And in terms of their performance and dignity, well, it's even worse. #Mahsa_Amini #Biological_Assassination**	0	1	1	1	Mix of Neutral and Hate (bold)
علیرضا دبیر: صحبت راجب سیاست شرعاً مشکل داره، از بروبچِ کشتی میخوام گوشیشون رو کنار بزارن و رو تمریناتشون تمرکز کنن. بعد از انقلاب اولین نفری که میدم سگم بهش تجاوز کنه تویی بی وجود# مهسا_امینی Alireza Dabir: "Talking about politics is religiously problematic. I ask the wrestling guys to put their phones aside and focus on their training." After the revolution, the first person I'll have my dog violate is you, you worthless being. #Mahsa_Amini	0	1	1	1	individual
مکالمه رعنا رحیم پور با مادرش چقدر مسخره و مصنوعی بود سریع هم اومد استوریش کرد که اره منم و جاسوسی بوده ، عرزشی ها هم بلافاصله عر عر کنان تجزیه تجربه میکنن خر همون مادرته جنده #مهسا_امینی Rana Rahimpour's conversation with her mother was so ridiculous and fake. She quickly posted it on her story, like 'Yeah, it's me, and it's been espionage.' The regime loyalists immediately started braying and analyzing it. The real fool here is your own mother, you whore. #Mahsa_Amini	0	1	1	1	Individual Politics
سه ساعته علاف این مامور های گازم اینم از برکات ج/ا که بعد ۴۳ سال تازهه شهر مارو که مرکز استان دارن لوله کشی گاز میکنن **پول خونم ازمون میگیرن** Three hours of gas workers, this is one of the blessings of the GCC. After 43 years, they are installing gas pipes in our city, the provincial capital**. They are charging us for our blood money**..	0	1	0	1	idiomatic
از ماست که بر ماست تا به این دین و باورهای بیابان گرد ملخ خوار باور داریم **همین آش و همین** کاسه It is up to us to believe in this religion and beliefs of the desert, the locust-eating locusts, **the same soup and the same bowl.**	0	1	0	1	idiomatic
از دی به سینا از سینا به نجمیه چچه غلطی دارند میکنند از نجمیه هم می‌خوان ببرند امیر علم حتما بیشرف ها From Day to Sina, from Sina to Najmieh—what the hell are they doing? Now they want to move from Najmieh to Amir Alam. They must be absolute scoundrels.	0	0	0	0	Background knowledge
هر افغانی چند ریاله؟ بستگی داره چند ساعت میخوای استفاده کنی How much is an Afghan worth? It depends on how many hours you want to use them for.	0	0	0	0	Background knowledge
قدرت یعنی عبا تنِ مسی کردن. نه گوزگوز Power means putting a cloak on Messi, not just **farting around..**	0	1	0	1	Background knowledge-idiom

Table 8: Transfer Learning Model Predictions on Tweets: Samples Illustrating Model Performance Across Different Experiments and Their Strength in Capturing Different Hate Types in Persian