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. Here we show how distant supervision can expand the available datasets for Persian (Farsi) while minimizing the dependence on manual labeling. With this enriched dataset, we assess the effectiveness of various large language models (LLMs) in detecting hate speech, vulgarity, and violent content in Persian. This establishes the first comprehensive benchmark for LLMs in Persian toxic language detection. As expected, these LLMs do not perform as well on Persian toxic detection as on English. We also consider the impact of cultural context on transfer learning for toxic content detection. Specifically, we show that languages with closer cultural similarities to Persian yield better results on transfer learning. Conversely, languages with more distinct cultural differences exhibit limited improvements. This underscores the critical role of cultural alignment in enhancing the performance of transfer learning models in this domain.

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 undermine the overall quality and inclusivity of online communities.

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 shown exceptional performance on a wide range of language-related tasks across multiple languages, there has been a growing interest in evaluating their effectiveness on toxic detection (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]. 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. Addressing the challenges of toxic language detection in Persian is critical, given its widespread use and, made more difficult by its non-Latin script, diverse writing styles, and regional dialects.

The current work is a comparative study on using different methods for Persian toxic-speech detection, including fine-tuning, data enrichment, zero-shot and few-shot of multiple LLMs, and transfer-learning across languages.

It aims to address 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. Could better performance be achieved using fine-tuning?

RQ3. Would data enrichment (using distant supervision) improve Persian toxic language detection?

RQ4. Given the fact that toxic speech classifiers are culturally insensitive (Lee et al., 2023), can transfer learning from particular languages enhance model performance? Which languages

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

lead to better performance?

We study these RQs through experiments on the PHATE dataset (Delbari et al., 2024), which covers three types of toxic language in Persian,- hate-speech, vulgarity, and violence. We find that toxic language identification in Persian continues to be a challenging task for most existing LLMs. However, tuning ParsBERT (Farahani et al., 2021) leads to better results, also outperforming other multilingual transformer-based models such as XLM-R and mT5. In addition, using distant supervision to obtain additional Persian training data, significantly enhances the performance of ParsBERT. We also find that transfer learning for toxic detection in Persian is highly dependent on cultural context. In particular, when there is a cultural overlap between the source and destination languages, the results tend to improve significantly.

2 Related Work

2.1 Toxic Language Detection

Early studies of toxic language detection focused on using Machine Learning (ML) and Deep Learning (DL) techniques for English hate speech detection 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). Similar efforts addressed offensive and abusive language detection (Bade et al., 2024; Aiyanyo et al., 2020; Cao et al., 2020; Risch et al., 2020), as well as violence and cyberbullying (Wang et al., 2020; Pamungkas and Patti, 2019; Van Hee et al., 2015; Guo and Gauch, 2024; Cano Basave et al., 2013; Huang et al., 2018).

Research has expanded to other languages, such as Indonesian (Ibrohim and Budi, 2019), Danish (Sigurbergsson and Derczynski, 2020), Arabic (Mubarak et al., 2021; Bensalem et al., 2023; Abuzayed and Elsayed, 2020), Korean (Jeong et al., 2022), Chinese (Deng et al., 2022), Greek (Pitenis et al., 2020), and Indic languages (Gupta et al., 2022), with notable studies on Hindi (Kapoor et al., 2019).

The emergence of LLMs has further advanced this field, with studies benchmarking their performance across various languages (Zampieri et al., 2020; Verma et al., 2022; Koufakou et al., 2020; Caselli et al., 2021; Saleh et al., 2023; Nguyen et al., 2023; Chiu et al., 2021; Zampieri et al., 2023). Shared tasks, such as 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 sparse. Existing studies (Jey et al., 2022; Sheykhlan et al., 2023; Safayani et al., 2024; Ataei et al., 2023; Delbari et al., 2024) provide limited publicly available datasets and primarily focus on a single category of toxic language. Notably, 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, the current work enhances the dataset with distant supervision, experiments with various LLMs and transfer learning techniques, and shifts 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 leverage 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 the transferability of language models.

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

The dataset PHATE (Delbari et al., 2024), which forms the foundation of our work, consists of 7,056 tweets distributed across four classes: 582 labeled as violence, 1,583 as vulgar, and 1,632 as hate. The remaining 3,259 tweets are 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 objective is not binary classification but rather distinct multi-class categorization, we dropped this overlapping label to focus on distinct toxic categories.

To apply distant supervision, we need to create a toxic lexicon for Persian. To build a toxic lexicon, we had three native Persian speakers carefully examine the dataset to identify frequently used keywords in each class. This initial examination resulted in 164 keywords, which we reduced to 127 by eliminating terms likely to be used in neutral contexts, such as specific names, to mitigate potential bias. The selection of these keywords was finalized using majority voting among the annotators. At this stage, nearly 40% of the keywords were related to vulgarity.

We then followed a structured approach for each toxic class to expand the lexicon further. 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, including racial and ethnic groups, religious groups, gender, individuals with disabilities, and other social groups (Silva et al., 2016). We further added another hate target, politics, as the frequency of this target in the dataset is high (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 a 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 further, we employed the FastText model (Bojanowski et al., 2017) trained on Persian to identify related and synonymous terms for the 377 keywords identified earlier. Filtering out duplicates and irrelevant words, produced 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.

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.

4 Experiments and Results

We use ParsBERT (Farahani et al., 2021) as our baseline model, as it is the only model exclusively pre-trained on Persian data, making it an essential benchmark for evaluating the performance of other multilingual models. Additionally, ParsBERT has

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

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)
Gemma 2	9B	(tea, 2024)
GPT 3.5 Turbo	175B	(Brown, 2020)

Table 1: LLMs used in our Study.

demonstrated promising results across a variety of Persian NLP tasks, further establishing its reliability and effectiveness for this domain. Table 1 provides the list of language models used in our benchmarking process. All models were trained for 10 epochs, and the final results on the test dataset are reported based on the epoch that achieved the highest F1 score on the validation set. This methodology ensures that we capture the optimal performance of each model during evaluation.

In our experiments, we fine-tuned different LLMs and evaluated their performances on both the enriched and baseline datasets to address two main objectives: (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 Persian dataset (PHATE) into English using the Google Translate API. We then evaluated Llama 3 on the translated dataset to further analyze its performance and the impact of language translation on classification results. This step underscores Llama 3's adaptability and robustness across different languages. Further elaboration on this will be provided in the discussion section.

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. Further, we employed GPT 3.5 Turbo in both zero-shot and few-shot settings to compare performance across each class. Additionally, we used 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 proven to yield 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.

Regarding transfer learning, we utilized three languages—Arabic, Indonesian, and English— and explored the interplay of linguistic and cultural factors in toxic speech detection. Since Llama 3 consistently achieved better results compared to other multilingual models, we selected this model for 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. Written in the Latin script, it is known for its straightforward grammar and simple phonetics. 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 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 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). To ensure comparability, we maintained fairly equal dataset sizes for all languages, with balanced label distribution across all classes. Notably, since we could

	Model	Violence			Hate			Vulgar			F_{macro}
		P	R	F_1	P	R	F_1	P	R	F_1	
Zero/Few shot	GPT 0-shot	35	75	48	39	$\underline{\mathbf{89}}$	54	61	**46**	52	51
	GPT 2-shot	40	$\underline{\mathbf{81}}$	54	55	69	61	79	37	50	55
	GPT 0-shot binary	$\underline{\mathbf{81}}$	73	$\underline{\mathbf{77}}$	**83**	64	72	$\underline{\mathbf{85}}$	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
	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 – Ar	-	-	-	75	$\underline{\mathbf{89}}$	81	81	$\underline{\mathbf{84}}$	82	82
	Llama 3 – Ar+Fa	-	-	-	86	88	$\underline{\mathbf{87}}$	**83**	$\underline{\mathbf{84}}$	$\underline{\mathbf{84}}$	$\underline{\mathbf{86}}$
	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 – Id	-	-	-	$\underline{\mathbf{89}}$	84	86	-	-	-	-
	Llama 3 – Id+Fa	-	-	-	85	83	84	-	-	-	-

Table 2: Toxic Detection Performance Across Different Approaches. The best performance for each group of approaches is presented in bold, and the overall best performance is underlined.

not find any dataset of vulgar or violent language with enough samples for training, we limited our Indonesian experiments to hate detection only.

4.1 Results

Table 2 presents a comprehensive comparison of model performance. 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.

4.1.1 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".

4.1.2 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, 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. Finally, Llama 3 – Instruct trained on the enriched dataset achieved consistent F1 scores of 63, 57, and 57 across the three categories.

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 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.

4.1.3 Transfer Learning

Since the results with the Llama 3-based model were better compared to other multilingual LLMs, we used this model for all transfer learning experiments in this study. 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.

Furthermore, fine-tuning on Arabic data alone (Llama 3 – Ar) resulted in strong F1 scores of 81 for both "Hate" and "Vulgar." Adding Persian data to the Arabic training set (Llama 3 – Ar + Fa) further enhanced performance, achieving the highest F1 scores of 87 for "Hate" and 84 for "Vulgar." This is the highest result among all experiments.

Regarding Indonesian, fine-tuning on this language alone (Llama 3 – Id) resulted in strong F1 scores of 86 for 'Hate.' However, adding Persian data to the Indonesian training set (Llama 3 – Id + Fa) decreased performance across all metrics, resulting in a slight drop in the F1 score to 84.

4.2 Analysis and Discussion

In this section, we address our research questions and provide some additional discussion.

4.2.1 RQ1: Generative LLMs Performance

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

Table 3 shows that, in zero/few-shot settings, GPT-3.5 Turbo demonstrated significantly better performance in binary classification tasks compared to multi-label classification. The model frequently mislabeled instances in zero-shot multi-label classification, particularly confusing labels such as 'hate' and 'violence.' Additionally, some instances of 'hate' are incorrectly classified as 'neutral.'

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. We observed that the model shows noticeably better performance, especially in violence detection, where the results even surpass those achieved through fine-tuning and transfer learning. After analyzing the errors in the binary setting, we found that GPT-3.5 Turbo 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 3 presents some misclassification samples by GPT 3.5 Turbo. Interestingly, the model's performance either remained steady or dropped as the number of shots increased. Ultimately, our analysis shows 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 tasks (Abaskohi et al., 2024).

4.2.2 RQ2: Fine-Tuning Effect

Could better performance be achieved using fine-tuning?

ParsBERT, among fine-tuned models, achieved a higher F-score across all classes. Despite being relatively smaller than other models, this monolingual model outperformed others significantly, highlighting the effectiveness of ParsBERT in handling Persian language tasks. However, in comparison to other reported tasks, (Farahani et al., 2021) ParsBERT still lagged in detecting toxic language.

While other models perform worse than ParsBERT, 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 observed no significant difference in performance. Using the translated dataset, we observed that all metrics dropped notably after this step. Upon examining the dataset, we found that the decline in performance stemmed from problematic translations, as most entries were informal and therefore, difficult for Google Translate to process correctly.

4.2.3 RQ3: Data Enrchiment via Distant Supervision

Would data enrichment (using distant supervision) improve Persian toxic language detection?

Our results demonstrate that distant supervision leads to improvement on mT5 and significant enhancement on BERT base models. However, it performs poorly on Llama 3 and Gemma 2. Notably, the metrics reveal that the results on Llama 3 are 50% worse than those on Gemma, 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.

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 2 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

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 3: Samples of misclassified instances in GPT-3.5 Turbo - English translation is literal

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
Ugh, die already! How much sugar did you add to it ای بمیری چقدر شکر بهش زدی	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 4: Samples (with translation) of misclassification instances after training ParsBERT on enriched dataset

"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 بمیر (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 4 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 4 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.

4.2.4 RQ4: Cross-Lingual Transfer Learning

Given the fact that toxic speech classifiers are culturally insensitive (Lee et al., 2023), can transfer learning across languages enhance model performance? Which languages lead to better performance?

To evaluate how well Persian can benefit from other languages, we experimented with three distinct languages: Arabic, English, and Indonesian. Our findings indicate that while Persian can effectively leverage the Arabic and Indonesian datasets, its performance gains from the English dataset are less pronounced. A 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. 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 allows Persian to exploit the morphological richness of Arabic during transfer learning, leveraging the ca-

Tweet	Actual Label	ParsBERT	Llama 3 base	TL Ar
مهریه اش رو جلوش میندازم ولی از حق مسلم طلاق و چندهمسری نمیگذرم. "I'll waive her dowry, but I won't give up the absolute right to divorce and polygamy.	Hate	Violence	Neutral	Hate
حق طلاق با مرده ولی زن با هرزگی واهرم مهریه هر کار بخواد میکنه قدری منصف باشید من عینا برایم اتفاق افتاد خیانت کرد تا مرد طلاقش بدهد و میلیاردها تومان مهریه بگیره. Divorce rights belong to the man, but women exploit promiscuity and the leverage of the dowry to do whatever they want. Be fair! This happened to me personally: she cheated to force the man to divorce her and took billions in dowry.	Hate	Neutral	Neutral	Hate
تف به مملکتی که دیه الناز رکابی از مهدی ترابی کمتره🤔😟!... Shame on a country where Elnaz Rekabi's diya (blood money) is less than Mehdi Torabi's... 😟🤔	Hate	Neutral	Vulgar	Hate
شاید تو کوچه ما عروسی نباشه ولی این عزا به کوچه شما هم میرسه Maybe there's no wedding on our street, but this mourning will reach your street too	Hate	Neutral	Neutral	Hate
هیچ چیز بهتر از این نیست که تیم مورد علاقت تیمای عربی رو له کنه وچیزی غم انگیز تر از این نیست که تیمت به عربا ببازه... heartbreaking than your team There's nothing better than your favorite team crushing Arab teams, and nothing more losing to the Arabs...	Hate	Neutral	Neutral	Hate
At the Riyadh concert, were locusts and camel urine also served? تو کنسرت ریاض ملخ و شاش شتر هم سرو میشد؟.	Hate	Neutral	Neutral	Hate

Table 5: Culturally-dependent hate instances detected via transfer learning (TL) from Arabic

pacity 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. This observation suggests that cultural influence may have a more significant impact than linguistic similarity.

Due to the lack of an Arabic and Indonesian dataset for the violence class, we limited our violence transfer experiments to English. Interestingly, these experiments demonstrated that English can still provide relevant contextual information about violence applicable to Persian.

To further explore the potential of transfer learning, we conducted supplementary experiments by integrating datasets from three language pairs (Arabic-Persian, English-Persian, and Indonesian-Persian). These experiments 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. However, notably, we observed a decline in all metrics with the combination of Indonesian and Persian. Further work will be needed to figure out why.

Since hate is the only class for which we found adequate data in all languages for our experiments, we present hate samples that were not classified correctly by Llama 3 and ParsBERT but were correctly predicted through the transfer learning setting in Table 5. This suggests that the proposed method provides sufficient contextual information for the model to detect this class accurately. Ultimately, as presented in Table 2, results from the integration of Arabic and Persian datasets yield higher results among all experiments.

5 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, which is currently the only monolingual Persian LLM. We also find that transfer learning yields better results when cultural similarities between languages are prioritized. Specifically, Persian benefits more from Arabic and Indonesian resources than from English, likely due to shared cultural contexts. This emphasizes the importance of considering cultural alignment when selecting source languages for transfer learning.

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.

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).

Abeer Abuzayed and Tamer Elsayed. 2020. Quick and simple approach for detecting hate speech in Arabic tweets. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 109–114, Marseille, France. European Language Resource Association.

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.

Qianjia Huang, Diana Inkpen, Jianhong Zhang, and David Van Bruwaene. 2018. Cyberbullying intervention based on convolutional neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 42–51, Santa Fe, New Mexico, USA. 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.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. 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.

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.

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.

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.

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

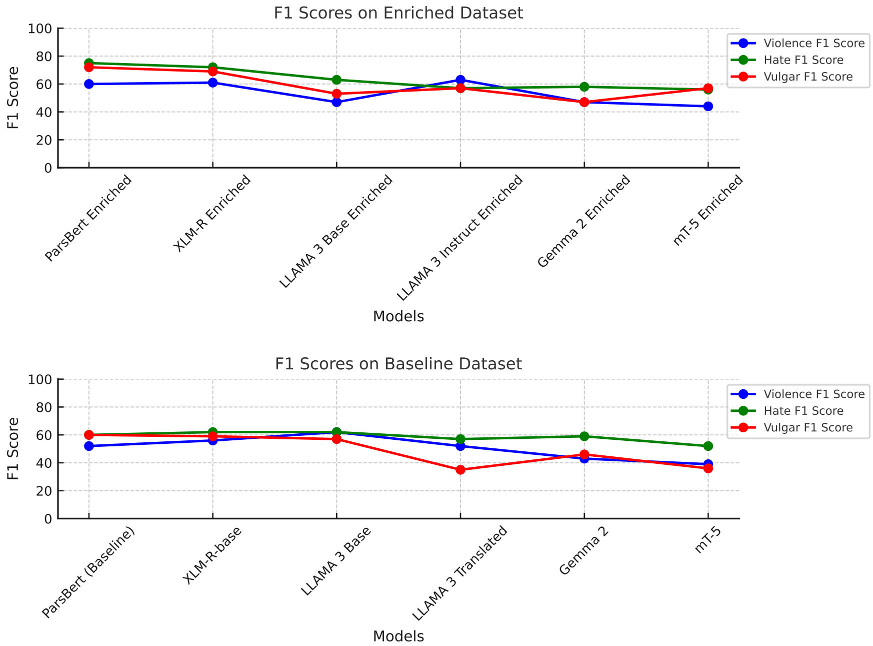

Figure 1: The fine-tuned models' performance before and after dataset enrichment.

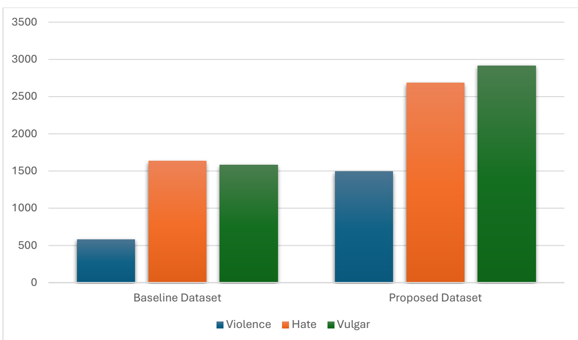

Figure 2: Label Distribution Before and After the Enrichment

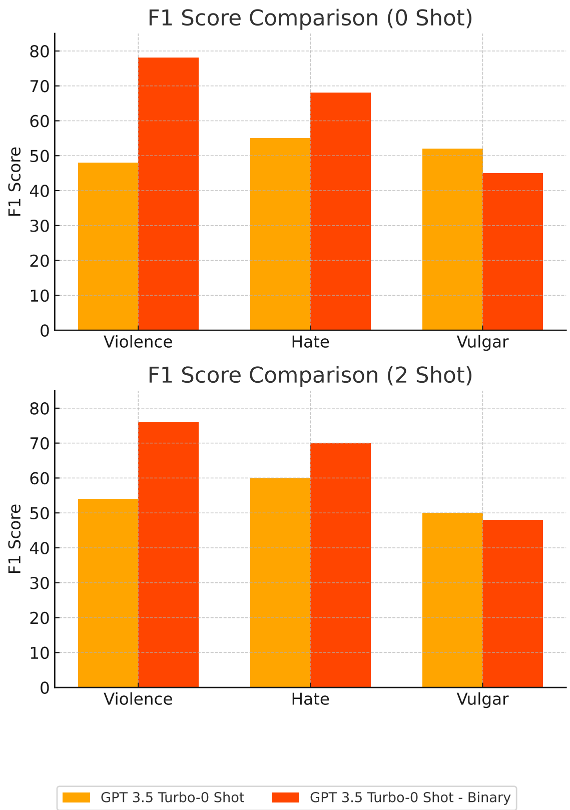

Figure 3: Comparison Between Binary Classification and Multi-Label Classification in 0-Shot and 2-Shot Configuration for GPT.

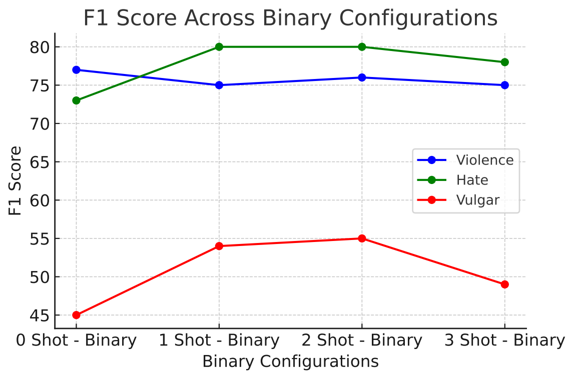

Figure 4: GPT F1 Score Results for Three Classes Across N-Shot Configurations

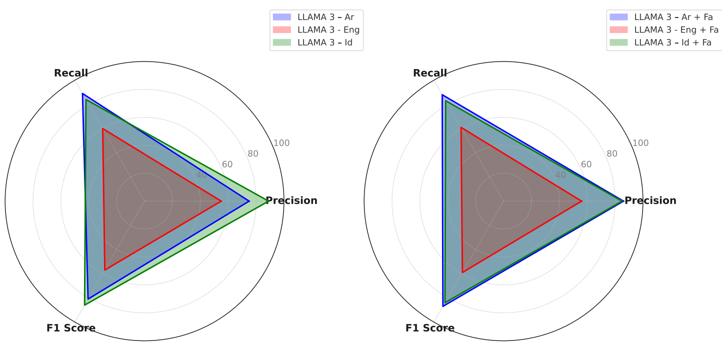

Figure 5: Performance of Transfer-Learning methods on the Test Set for Hate Speech Detection.

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 6: The Prompt Used for the GPT Experiment