Exploring Subjective Tasks in Farsi: A Survey Analysis and Evaluation of Language Models

Donya Rooein[1], Flor Miriam Plaza-del-Arco[1,2], Debora Nozza[1] , Dirk Hovy[1]
[1]Bocconi University, Milan, Italy
[2]Leiden University
Correspondence: donya.rooein@unibocconi.it

Abstract

Given Farsi's speaker base of over 127 million people and the growing availability of digital text, including more than 1.3 million articles on Wikipedia, it is considered a "middle-resource" language. However, this label quickly crumbles when the situation is examined more closely. We focus on three subjective tasks (Sentiment Analysis, Emotion Analysis, and Toxicity Detection) and identify significant challenges in data availability and quality, despite overall increases in data availability. We review 110 publications on subjective tasks in Farsi and observe a lack of publicly available datasets. Furthermore, existing datasets often lack essential demographic factors, such as age and gender, that are crucial for accurately modeling subjectivity in language. When evaluating prediction models using the few available datasets, the results are highly unstable across both datasets and models. Our findings show that the volume of data alone is insufficient to improve a language's standing in NLP.

1 Introduction

Many NLP tasks, like emotion classification, are inherently subjective. There are different valid perspectives on the "correct" data labels. How emotions are perceived, for example, differs between the sender and the receiver's subjective interpretations (Barz et al., 2025). The same message, expressing frustration or sarcasm, could be interpreted humorously by one individual but taken offensively or negatively by another, depending on their cultural background, personal experiences, or situational context.

Subjective tasks in NLP, such as emotion analysis, sentiment analysis, and toxic detection, have received increasing attention because they directly affect societal aspects, including decision-making, customer feedback, product evaluation, and the understanding of social dynamics in general (Nandwani and Verma, 2021). These tasks involve assigning texts to specific emotions or sentiments that best reflect the author's mental or emotional state (Tao and Fang, 2020). Recent surveys in emotion and sentiment analysis (Murthy and Kumar, 2021; Kusal et al., 2022; Singh Tomar et al., 2023; Hung and Alias, 2023; Venkit et al., 2023; Al Maruf et al., 2024; Plaza-del Arco et al., 2024; Song et al., 2025) have primarily focused on identifying available datasets, reviewing models, exploring detection techniques across various modalities (e.g., visual, vocal, textual), and discussing applications. These studies focus on English and do not consider other languages such as Farsi[1].

Language technologies play a crucial role in promoting multilingualism and preserving linguistic diversity worldwide. However, many languages still face challenges in resource availability, particularly for subjective tasks, despite having substantial digital resources and peer-reviewed research. This is the case for Farsi, which has over 1.3 million Wikipedia articles[2] and has been classified by Joshi et al. (2020) as a language with a strong web presence but insufficient efforts in labeled data collection, ranking just below high-resource languages. Despite these resources, **research on subjective tasks in Farsi remains notably scarce**, making it a low-resource language in this domain.

The Farsi script, also known as the Perso-Arabic script, is a modified form of the Arabic alphabet (Izadi et al., 2006). Persian and its dialects are official languages in Iran, Afghanistan, and Tajikistan, and are also spoken by minority communities in Uzbekistan, Turkmenistan, Azerbaijan, Armenia, Georgia, and southern

[1]Also known as Persian.
[2]https://en.wikipedia.org/wiki/Persian_Wikipedia

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 83–95

Russia. While Farsi and Arabic scripts are often considered similar due to their shared roots, they differ in their alphabets and their writing styles (Izadi et al., 2006).

While a few survey studies in Farsi focus on sentiment analysis and discuss resource limitations and methodological developments (Rajabi and Valavi, 2021; Asgarnezhad and Monadjemi, 2021; Borowczyk, 2023), to the best of our knowledge, no existing work provides a comprehensive survey of multiple subjective tasks in Farsi. The scope of this Paper includes evaluating different encoder-only models and LLMs across three key tasks: emotion analysis (EA), sentiment analysis (SA), and toxic detection (TD). These tasks span a range of applications and research communities, from traditional sentiment analysis of product reviews to offensive-language detection. This list is by no means exhaustive for subjective language tasks such as humor detection, sarcasm detection, and metaphor detection. Rather, we highlight the areas with the most dataset availability in the literature we surveyed. This gap is particularly concerning in the era of LLMs, where these systems are not only widely accessible but also increasingly used for subjective discussions (Ouyang et al., 2023). It is essential to evaluate their ability to understand and process sentiment and emotion in Farsi, as well as to assess their handling of toxicity to ensure safe and responsible interactions. The lack of research in this area underscores the urgent need for a focused exploration to ensure that Farsi, like other languages, benefits from advancements in subjective NLP.

We collect relevant studies from publications drawn primarily from ACL Anthology[3], and complemented by additional searches on Google Scholar[4]. We report the available dataset for each task, including important metadata such as dataset size, labels, and source. Additionally, we evaluate various language models on selected datasets to assess their capabilities for these subjective tasks in Farsi.

We present the following key contributions:

- A detailed survey of publications, datasets, and resources specific to the three subjective tasks in Farsi: sentiment analysis, emotion analysis, and toxicity detection.
- An experimental evaluation of encoder-only multilingual models and open-source LLMs on these tasks in Farsi.
- An analysis of the impact of text translation as a potential solution to address low-resource challenges in Farsi language.

2 Background

Subjective tasks such as EA, SA, and TD often pose unique challenges due to their reliance on context, cultural nuances, and linguistic features. The EA involves classifying emotions expressed in a text (e.g., joy, sadness, anger) (Alm et al., 2005). For instance, recognizing the nuanced difference between Farsi expressions like "دلش گرفت" (delash gereft, literally "his/her heart became tight") conveying sadness, versus "دلشوره دارد" (delshooreh dārad, literally "he/she has a salty heart") depicting anxiety, requires deep cultural and contextual understanding compared to relatively straightforward English expressions like "feeling sad" or "feeling anxious". The SA consists of determining the sentiment polarity of a text, typically positive, negative, or neutral (Wilson et al., 2005). For example, the Persian expression "جای تو خالیه" (jāye to khālie, literally "your place is empty") carries a positive sentiment, often implying affection, inclusion, and the speaker expresses a desire for the listener's presence. However, translated directly into English, it may suggest loneliness, absence, or even negativity. Such examples underscore the importance of accurately capturing sentiment, which requires sensitivity to cultural context and linguistic nuances. Toxicity detection consists of identifying language or content considered harmful, offensive, abusive, hateful, or otherwise inappropriate (Pavlopoulos et al., 2020). The interpretation of what constitutes toxic content often varies significantly based on cultural and societal norms. For example, the phrase "کمه عقلت" ("you're not very smart") in Farsi might be considered mildly humorous among close friends but is perceived as offensive in formal or public contexts.

3 A Survey on NLP Studies Covering Subjective Tasks in Farsi

To identify relevant papers with resources related to For EA, SA, and TD tasks in Farsi, we design a structured search query comprising

[3] https://aclanthology.org/anthology+abstracts.bib

[4] https://scholar.google.com/

three main components: <Task>, <Dataset>, and <Language>[5]. The <Task> component includes the three NLP tasks we explore: the EA, SA, and TD. To ensure a comprehensive selection of studies for these tasks, we identify papers whose titles or abstracts include keywords associated with each task. For the EA task, our query includes the terms "emotion classification", "emotion detection", "emotion recognition", "emotion analysis", and "emotion prediction". For the SA task, we incorporate the following keywords: "polarity classification", "sentiment classification", and "sentiment analysis". Lastly, for the TD task, we use terms including "hate speech detection", "offensive language detection", "hate speech classification", "offensive language classification", "toxic detection", and "toxic classification". The <Dataset> component includes related terms, i.e., "data set," "dataset," "corpus", and "corpora". Finally, the <Language> component explicitly focuses on language-related terms, namely "Farsi" and "Persian". Our query variations are derived from 5 keywords associated with the EA, 3 with the SA, and 6 with the TD tasks (a total of 14 keywords), combined with 4 dataset formulation strategies and 2 for the language, yielding a final total of 112 unique phrase searches. To further expand our search, we also collect publications using only <Task> and <Language>, adding 28 additional search phrases. In total, we executed 140 unique phrase searches.

We identify 12 unique papers from the ACL Anthology: eight focused on SA, four on EA, and none on the TD task. This absence indicates the lack of research and publicly available datasets on Farsi toxicity detection in the ACL Anthology. To expand our search results, we also use Google Scholar. Google Scholar lists papers from different research databases; however, it is difficult to verify all the returned sources. We use the SerpApi[6] library to retrieve papers from Google Scholar. To limit the search results from this engine, we configure the SerpApi to only return the top 10 relevant papers for a given search keyword. This limitation allows us to verify their publishers manually. This search strategy adds 98 more papers which 40 from arXiv[7], 16 from IEEE[8], 12 from Springer[9], and 30 from other publishers.

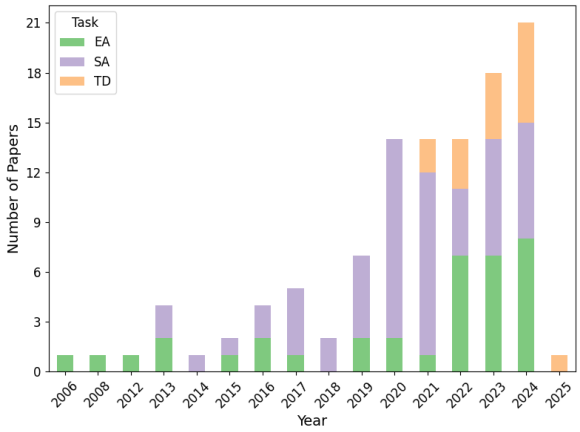

Figure 1: Distribution of papers considered in our survey by year and tasks (EA: Emotion Analysis, SA: Sentiment Analysis, and TD: Toxicity Detection).

Thus, we have a total of 110 papers[10](only 11% from ACL Anthology), with 36 papers for EA, 58 papers for SA, and 16 papers for TD[11]. Figure 1 shows the statistics of the collected papers published from 2006 to 2025. The SA task represents the largest share at 52.7% (58 out of 110) of all papers, followed by EA at 32.7% (36 out of 110). The EA task began among the non-NLP community in 2006, focusing on EA through speech. The number of publications remained low in the early years; however, by 2024, the EA in Farsi had increased to 8 papers incorporating text-based modalities. The TD task, which did not appear until 2021, already accounts for nearly 14.5% (16 out of 110) of papers by 2025, indicating that TD is becoming an increasingly important area of research in NLP Farsi.

3.1 Annotation Criteria

After identifying relevant papers, we conduct a manual annotation to summarize and categorize the papers based on consistent criteria. The motivation here is to identify publicly available datasets in Farsi for each task. We adopt the annotation framework proposed by Plaza-del Arco et al. (2024), which suggests surveying EA datasets based on five key aspects: annotation framework, language, multimodal, content source, and dataset size. We expand this framework

[5]All searches are updated by March 2025.
[6]https://serpapi.com/
[7]https://arxiv.org/
[8]https://www.ieee.org/
[9]https://www.springer.com/
[10]The list of the reviewed papers is available at https://anonymous.4open.science/r/subjective_tasks_farsi-8874/README.md
[11]Three of these papers are in the Farsi language and were published at local conferences within Iran.

to all the considered subjective tasks and include additional details: lexicon, the type of classification task (e.g., binary, multiclass, or multilabel), and, specifically for studies involving dataset creation, whether the demographics of annotators are explicitly considered. We also include information on the availability of datasets used in each paper.

Our annotation results reveal several trends. For the data modalities, most works (86.4%) are text-based, a few (4.5%) combine text with speech, and 8.2% focus on speech-only datasets. In addition, only one paper (0.9%) uses acoustic and visual data. The datasets used in the reviewed papers are from social media platforms, e-commerce websites, and specialized corpora. The most frequently used sources for social media for all tasks are X[12] (previously Twitter) and Instagram[13].The e-commerce source is mostly Digikala[14], Iran's largest online retail platform, which contains extensive user-generated product reviews that are valuable for sentiment analysis. Additional sources include datasets from Booking.ir[15], a popular platform for hotel reviews, movie review comments[16]. In some cases, authors use specialized resources such as radio plays or collect datasets from surveys of specific populations. Based on our review collection for data availability, we identify three categories of papers: (I) papers without datasets, (II) papers with datasets that are not publicly available, and (III) papers with publicly available datasets. We identify 17 of the 36 EA papers as dataset papers, but only 7 provide publicly available datasets. In particular, 4 of these 7 datasets are from the ACL Anthology. For SA, we identify 33 dataset papers, but only 5 datasets are available (3 from the ACL Anthology). Finally, TD has 14 papers on datasets, 3 of which are publicly available, and none from the ACL Anthology. In total, we therefore identify 15 publicly available datasets across the three tasks. In the following sections, we provide further details on the available datasets for each task, along with their characteristics.

[12] https://x.com/
[13] https://www.instagram.com/
[14] https://www.digikala.com/
[15] https://www.booking.ir/
[16] From https://cinematicket.org/

3.2 Datasets

Table 1 presents a list of publicly available datasets along with detailed information on their names, label sources, data sources, sizes, and modalities.

EA datasets: We identify seven datasets for EA. The **Shemo** (Yazdani et al., 2021) dataset is derived from radio plays and annotates five primary emotions, i.e., anger, fear, happiness, sadness, and surprise along with a neutral category, comprising 3,000 samples. This dataset is the only dataset with both text and speech modality, and the rest of the datasets are text-only. **ShortPersianEmo** (Sadeghi et al., 2021) is from comments on the Digikala website, an e-commerce platform in Iran. The **SAT** (Elahimanesh et al., 2023) dataset originates from chatbot conversations and distinguishes a broader spectrum of emotions (happy, angry, anxious, ashamed, disappointed, disgusted, envious, guilty, insecure, loving, sad, and jealous) across 5,600 samples. The SAT dataset also includes the demographic information (age and gender) of participants. **ArmanEmo** (Mirzaee et al., 2022) and **LetHerLearn** (Hussiny and Øvrelid, 2023), **EmoPars** (Sabri et al., 2021a) consist of tweets annotated with common emotions such as anger, fear, sadness, happiness, and either wonder or surprise. In particular, EmoPars is annotated by a multilabel annotation approach, assigning a numerical value between 0 and 5 to each emotion (anger, fear, happiness, hatred, sadness, and wonder). None of these datasets fully adhere to well-known frameworks for emotion analysis such as Ekman's framework (Ekman et al., 1999) which includes anger, fear, sadness, joy, disgust, and surprise or Plutchik's model (Plutchik, 1982), which encompasses eight primary emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. **LearnArmanEmo** (Hussiny et al., 2024) combines ArmanEmo and LetHerLearn by unifying their labels based on Ekman's framework. In this unified approach, the label "happiness" is used instead of "joy", and an "other" category is added to capture emotions outside the defined set.

SA datasets: Pars-ABSA (Shangipour ataei et al., 2022), Persian Digikala (Kobari et al., 2023), and Persian-English code-mixed datasets (Sabri et al., 2021b) categorize sentiment of Farsi sentences into positive, negative, and neutral

Task	Dataset	Mult.	Labels	Source	Size	Included
EA	Shemo	T, S	E + [neutral]	radio plays	3,000	-
EA	ShortPersianEmo	T	[happiness, sadness, anger, fear, other]	e-commerce	5,472	-
EA	SAT	T	E + [anxious, ashamed, disappointed, envious, guilty, insecure, loving, jealous]	chatbot conv.	5,600	-
EA	ArmanEmo	T	E - [disguss] + [hate, other]	social media	7,000	✓
EA	LetHerLearn	T	E + [other]	social media	7,600	✓
EA	LearnArmanEmo	T	E + [other]	social media	14,880	-
EA	EmoPars	T	E - [disgust] + [hatred]	social media	30,000	✓
SA	SentiPers	T	[−2, −1, 0, +1, +2]	e-commerce	15,683	✓
SA	Pers.-Eng.	T	[negative, neutral, positive]	social media	3,640	-
SA	Persian Digikala	T	[negative, neutral, positive]	e-commerce	34,465	-
SA	Pars-ABSA	T	[negative, neutral, positive]	e-commerce	10,002	✓
SA	MirasOpinion	T	[-1, 0, +1]	e-commerce	93,868	✓
TD	Phate	T	[hateful (violence, hate, vulgar), normal]	social media	7,056	✓
TD	PHICAD	T	[hate/offense, obscene, spam, none]	social media	300,000	✓
TD	Pars-OFF	T	[offensive, not-offensive]	social media	8,334	✓

Table 1: Overview of publicly available and private datasets used for subjective tasks in Farsi. **Task** presents Emotion Analysis (EA), Sentiment Analysis (SA), and Toxicity Detection (TD). The columns provide details on the dataset name if provided (**Dataset**), which content modality that dataset uses (**Mult.**), annotation labels (**Labels**), source of the data (**Source**), the dataset size (**Size**), and if they are included in our experiments (**Included**). [E] Ekman framework. [T] Text and [S] Speech.

Task	Prompt
EA	Given a text, identify the main emotion expressed. You have to pick one of the following emotions: [List of dataset labels]. Text: {input} Only answer with the emotion and omit explanations. Emotion: {output}
SA	Given a text, identify the sentiment expressed. You have to pick one of the following sentiments: [List of dataset labels]. Text: {input} Only answer with the sentiment and omit explanations. Sentiment: {output}
TD	Does the following text contain [hate speech/ offensive language]? Only answer with yes or no. Text: {input}, Hate speech: {output}

Table 2: Prompt templates for Emotion Analysis (EA), Sentiment Analysis (SA), and Toxicity Detection (TD) tasks.

labels. In particular Persian-English code-mixed dataset provides 3,640 labeled tweets, making it one of the few resources addressing sentiment in code-mixed Persian-English text. **SentiPers** (Hosseini et al., 2018) contains 15,683 Digikala reviews annotated on a five-point scale ranging from -2 to $+2$. **MirasOpinion** is the largest available dataset collected from Digikala for SA in Farsi language with 93,868 samples. They label each sentence by using a Telegram[17] bot to several users. They ask them to label the represented document as positive, negative, or neutral.

TD datasets: We find three datasets, each exclusively in text. **Phate** (Delbari et al., 2024) contains tweets that distinguishes between hateful content (with subcategories of violence, hate, and vulgar) and normal content, comprising 7,056 samples. The **PHICAD** (Davardoust et al., 2024) dataset is significantly larger, containing 300,000 samples, and labels content into hate/offense, obscene, spam, or none, also sourced from comments on the Instagram platform. Lastly, **Pars-OFF** (Ataei et al., 2023) focuses on a binary classification of offensive versus non-offensive content with 8,334 samples of tweets.

These datasets, while valuable for advancing Farsi subjective analysis tasks, face several limitations. Many of them exhibit a narrow focus in terms of data sources, mostly based on tweets and comments on the Digikala platform, which may limit the generalizability of models trained on them to other contexts. Moreover, they also suffer from a lack of demographic information. Only two datasets of EA (Shemo and SAT) provide the demographic factors (e.g., gender in Shemo and age and gender for SAT). Only authors of three datasets (Yazdani and Shekofteh, 2022) provide detailed documentation on how annotations were conducted, whether multiple annotators were used, or what inter-annotator agreement was achieved. Without such information, it is difficult to assess the reliability of the labels used to train or evaluate

[17] https://web.telegram.org/

models.

Evaluating these datasets using LLMs may help address some of these shortcomings. Abaskohi et al. (2024) shows the low performance of GPT3.5 and GPT4[18] on the emotion recognition task using only the ArmanEmo dataset. In the following section, we extend these evaluations by using various open-source models and datasets.

4 Evaluation Setting

4.1 Data

To evaluate the performance of language models on subjective tasks in Farsi, we select three datasets per task. For EA, we use ArmanEmo, LetHerLearn, and EmoPars. Since EmoPars contains multilabel emotions, we filter the dataset to include only samples in which one emotion has a non-zero value while all others are zero. With this approach, we reduce the size of the EmoPars dataset to 5,226 samples. We exclude the Shemo dataset because it relies on speech data, and the transcriptions alone do not adequately capture the nuances of emotion. We also excluded the SAT dataset due to its large number of labels, which could negatively affect language model performance. Finally, we exclude the LearnArmanEmo dataset, as it is derived from the LetherLearn and ArmanEmo datasets. For SA, we use ParsABSA, SentiPers, and a subsample of MirasOpinion. Since MirasOpinion is a very large dataset, we evaluate our language models on 30k randomly selected samples. We exclude the Persian-English code-mixed dataset due to its limited size and its primary focus on Persian code-mixed vocabulary. For the TD tasks, we use all the available datasets presented in Table 1. Given that the PHICAD dataset is extensive, with 300,000 samples, we experiment on a subsample provided by Davardoust et al. (2024)[19]of the dataset with 131,959 instances.

4.2 Models

4.2.1 Open Source Decoder-only Models

From the family of decoder-only LLMs, we select three instruction-tuned versions of popular open-source models, which are Meta-Llama-3-8B-Instruct (Dubey et al., 2024), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), and Qwen2-7B-Instruct (Yang et al., 2024). For each task, we use a zero-shot approach to detect the relevant labels of emotions for EA, sentiments for SA, and hate speech/offensiveness for TD. We use two different prompting strategies on a subset of EA and SA datasets (see Appendix B), then we use the following prompt template that yielded the best performance across these datasets. For TD, we exclusively use the prompt introduced by Delbari et al. (2024). We summarize the list of prompts in Table 2. For the EA and SA template, we ask the model to identify the main emotion and sentiment expressed in the text, selecting from a predefined list of dataset-specific labels.

Task	Model	Lang.	Template		Avg. F1
			(I)	(II)	
EA	Llama3-8B	FA	0.19	0.19	0.19
		EN	0.18	0.20	0.19
	Mixtral-7B	FA	0.20	0.19	0.19
		EN	0.20	0.19	0.19
	Qwen2-7B	FA	0.19	0.20	0.19
		EN	0.20	0.17	0.18
SA	Llama3-8B	FA	0.46	0.64	0.55
		EN	0.46	0.48	0.47
	Mixtral-7B	FA	0.50	0.77	0.63
		EN	0.48	0.54	0.51
	Qwen2-7B	FA	0.48	0.36	0.42
		EN	0.48	0.46	0.47

Table 3: The performances of LLMs in macro average F1 scores for two prompting templates on the EA task for the EmoPars dataset and SA for the MirasOpinion dataset are reported. We use Farsi (FA) and English (EN) versions of datasets (Lang.). The EN version is translated by the NLBB model. Avg. F1 is the average F1 score across templates.

4.2.2 Data Translation Experiments

Etxaniz et al. (2024) suggest that translating non-English datasets to English can enhance the performance of multilingual LLMs. We adopt this strategy by translating our datasets to assess their impact on model results. Since multiple machine translation systems are available, we first translated a subsample of 100 Farsi sentences using Google Translate[20], the NLBB model (Costa-Jussà et al., 2022), and GPT-4o. After manual evaluation, we found that Google Translate produced the lowest-quality translations. Both NLBB and GPT-4o provided acceptable results, though they still exhibited issues such as literal translations, mistranslations, and omissions.

[18] https://openai.com/

[19] Part 1 available at https://github.com/davardoust/PHICAD

[20] https://translate.google.com/

Ultimately, we chose to use NLBB because it is open source.

4.2.3 Encoder-only Models

For encoder-only architectures, we adopt standard fine-tuning approaches using XLM-RoBERTa (Conneau et al., 2020) and FaBERT (Masumi et al., 2025). XLM-RoBERTa is a multilingual transformer-based language model pre-trained on data from more than 100 languages. FaBERT is a Persian BERT base model pre-trained on the HmBlogs corpus, which includes both informal and formal Farsi texts. We fine-tune both models on nine datasets spanning the EA, SA, and TD tasks. Fine-tuning is performed by adding a classification head on top of the model's final hidden representations and optimizing it using a cross-entropy loss.

5 Results

In this section, we present the outcomes of our experiments, detailing the evaluation of prompt selection, LLMs' performances on the datasets in Farsi and their translation in English, and the fine-tuning approach.

5.1 Experiment 1: Prompt Variations and Data Translation

Prompt variations, even the smallest of perturbations such as adding a space at the end of a prompt, can affect the LLM's output (Salinas and Morstatter, 2024). In this regard, we include two prompting strategies: the first involves directly asking the LLM to identify the subjective label of a given text, while the second includes the text's data source as part of the prompt. For EA, we use a subsample from the EmoPars dataset, and for SA, we select the subsample of the MirasOpinion dataset. We choose these two publicly available datasets, because they are from the ACL Anthology and they have the largest sample sizes, with sample sizes of 5,226 for EmoPars and 30k for MirasOpinion. We evaluate two distinct prompt templates, as described in Appendix B, on these sub-samples.

Table 3 shows the performance of selected LLMs in EA and SA tasks over selected sub-samples in Farsi and English. The results of EA exhibit low F1-scores (between 0.18–0.20) across all models and configurations, with minimal differences between the original (FA) and translated (EN) data and only marginal variations due to template changes. Using English translation does not consistently improve the results. In the EA, translation to English has a minimal overall impact, with two models showing no change (Llama3-8B and Mixtral-7B). For Qwen2-7B, we observe a slight decrease in the English version of the data. The same trend is for the SA task, where all models have a lower average F1 score over English texts, except for the Qwen2-7B model, whose translation increases the average F1-score from 0.42 to 0.47, which is negligible. Regarding different prompt templates, we do not observe significant improvements over a specific template in the EA task. However, in the SA task, template (II) performs better than both the Farsi and English versions of the data, except for the Qwen2-7B model. These findings suggest that both prompt design and data translation strategies for these subjective tasks in the Farsi language have a slight influence on model outcomes, particularly in EA.

5.2 Experiment 2: LLM and Fine-Tuned LM Evaluations

Based on results in Table 3, we use the prompt template (II) and datasets in Farsi (no translation) in a zero-shot setup to evaluate the different LLMs' performances across the selected datasets in Section 4.1. Table 4 presents the macro average F1-score, across all tasks, datasets, and LMs. Performance is benchmarked against two baselines: a random classifier and a Most Frequent Class (MFC) baseline.

First, across all tasks, the fine-tuned transformer encoders (XLM-RoBERTa and FaBERT) consistently outperform the LLMs and baselines. This performance gap is most pronounced in Emotion Analysis, where XLM-RoBERTa and FaBERT achieve average F1-scores of 0.554 and 0.586, respectively, compared to 0.332–0.370 for the LLMs. Across all LLMs, Qwen2-7B consistently outperforms the other models, achieving the highest average F1-score in EA (0.370), SA (0.563), and TD (0.809). At the dataset level, Qwen2-7B achieves higher scores than ArmanEmo and LetHerLearn. Over the EmoPars dataset, Llama3-8B achieves a 0.227 average F1-score, which is slightly better than Qwen2-7B's 0.218 average F1-score. In the SA task, the differences between model classes narrow but remain significant. Qwen2-7B achieves the strongest average among LLMs (0.563), outperforming both Llama3-8B (0.534)

		LLMs			Fine-tuned LMs		Baselines	
Task	**Dataset**	**Llama3-8B**	**Mixtral-7B**	**Qwen2-7B**	**XLM-RoBERTa**	**FaBERT**	**Random**	**MFC**
EA	ArmanEmo	0.426	0.296	0.510	0.630	0.641	0.135	0.061
	LetHerLearn	0.343	0.348	0.383	0.653	0.550	0.151	0.048
	EmoPars	0.227	0.188	0.218	0.380	0.566	0.152	0.048
	Avg.	0.332	0.277	0.370	0.554	**0.586**	0.146	0.052
SA	ParsABSA	0.501	0.498	0.444	0.856	0.441	0.242	0.168
	SentiPars	0.453	0.351	0.562	0.564	0.570	0.199	0.108
	MirasOpinion	0.647	0.608	0.683	0.854	0.854	0.330	0.230
	Avg.	0.534	0.486	0.563	**0.758**	0.622	0.257	0.169
TD	Phate	0.674	0.682	0.562	0.748	0.849	0.504	0.412
	Pars-OFF	0.696	0.741	0.925	0.854	0.889	0.491	0.412
	PHICAD	0.770	0.789	0.942	0.950	0.950	0.500	0.418
	Avg.	0.640	0.737	0.809	0.851	**0.896**	0.499	0.414

Table 4: Macro average F1-scores for each model and dataset across three tasks: SA = Sentiment Analysis, TD = Toxicity Detection, EA = Emotion Analysis. Averages are calculated per task. MFC is Most Frequent Class. The highest average F1-score per task is highlighted in bold.

and Mixtral-7B (0.486). This suggests that Qwen2-7B may encode sentiment-related features more effectively. However, the fine-tuned models again show superior performance. XLM-RoBERTa, in particular, achieves an average score of 0.758, an improvement over Qwen2-7B. Interestingly, FaBERT performs unevenly across datasets: it delivers competitive or even equal performance on MirasOpinion and SentiPars but falls behind on ParsABSA. This inconsistency may indicate dataset-specific biases or annotation standards that affect the model's generalizability.

In the TD task, where all model families achieve their highest overall performance, LLMs approach the fine-tuned models. Qwen2-7B performs remarkably well, achieving an average score of 0.809, which is close to XLM-RoBERTa (0.851) and not far behind FaBERT (0.896). This strong performance aligns with the observation that toxicity classification relies heavily on identifying explicit lexical cues and linguistic markers that may be well represented in multilingual pretraining corpora. In this task, Qwen2-7B not only outperforms the other LLMs but also surpasses XLM-RoBERTa on Pars-OFF and PHICAD, suggesting that Qwen2-7B may be particularly well equipped for tasks involving offensive or toxic language categorization.

Taken together, these findings reveal that supervised fine-tuning on task-specific data remains essential to achieve better performance on these subjective tasks in Farsi, even with increasingly powerful multilingual LLMs. In addition, although zero-shot LLMs do not yet match the performance of fine-tuned encoder-based models, their relative strength in the TD task suggests that some subjective tasks may be more amenable to zero-shot inference than others. We also report the results per label for each dataset, task, and model in Tables 5 to 7 at the Appendices B.5 to B.7.

6 Conclusion

Research on subjective tasks in Farsi has grown over the past five years, with a notable increase in SA and TD research starting in early 2020. Most work has focused on two main data sources: social media data, such as tweets, and e-commerce data from Digikala, highlighting the scarcity of Farsi-language data sources. We reviewed over 110 papers, including 12 from the ACL Anthology and 98 from other publishers. We identified several gaps in these studies, including a lack of diverse datasets, annotation information, and demographic features in subjective tasks, particularly for EA. These gaps include demographic disparities such as age and gender and a lack of interdisciplinary research. Our experiments indicate that LLMs perform relatively poorly on EA tasks in Farsi but perform better on SA and TD. Additionally, fine-tuning consistently improves performance across all tasks.

7 Limitations and Ethical Considerations

We acknowledge several limitations in our study. First, our evaluation relies heavily on existing publicly available datasets, which may not comprehensively capture the linguistic, cultural, or topical diversity of the Farsi language. These datasets may contain annotation biases, domain-specific skew, or inconsistencies that could affect model performance and generalizability. Moreover, we use machine translation for English versions of Farsi texts, which may introduce semantic drift or cultural misrepresentation and affect fairness and accuracy. Another limitation is the limited number of models that support Farsi.

Acknowledgment

Donya Rooein, Debora Nozza, and Dirk Hovy are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Donya Rooein and Dirk Hovy's research is supported through the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). Debora Nozza's research is from the ERC under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE).

References

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.

Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: a survey. *IEEE access*, 12:18416–18450.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Razieh Asgarnezhad and S Amirhassan Monadjemi. 2021. Persian sentiment analysis: feature engineering, datasets, and challenges. *Journal of applied intelligent systems & information sciences*, 2(2):1–21.

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.

Christina Barz, Melanie Siegel, Daniel Hanss, and Michael Wiegand. 2025. Understanding disagreement: An annotation study of sentiment and emotional language in environmental communication. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 1–20.

Magdalena Borowczyk. 2023. *1 Research in Persian Natural Language Processing – History and State of the Art*, pages 1–24. De Gruyter Mouton, Berlin, Boston.

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

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

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

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*.

Paul Ekman, Tim Dalgleish, and M Power. 1999. Basic emotions. *San Francisco, USA*.

Sina Elahimanesh, Shayan Salehi, Sara Zahedi Movahed, Lisa Alazraki, Ruoyu Hu, and Abbas

Edalat. 2023. From words and exercises to wellness: Farsi chatbot for self-attachment technique. *arXiv preprint arXiv:2310.09362*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: a sentiment analysis corpus for persian. *arXiv preprint arXiv:1801.07737*.

Lai Po Hung and Suraya Alias. 2023. Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 27(1):84–95.

Mohammad Ali Hussiny and Lilja Øvrelid. 2023. Emotion analysis of tweets banning education in Afghanistan. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 271–277, Toronto, Canada. Association for Computational Linguistics.

Mohammad Ali Hussiny, Mohammad Arif Payenda, and Lilja Øvrelid. 2024. PersianEmo: Enhancing Farsi-Dari emotion analysis with a hybrid transformer and recurrent neural network model. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 257–263, Torino, Italia. ELRA and ICCL.

Sara Izadi, Javad Sadri, Farshid Solimanpour, and Ching Y Suen. 2006. A review on persian script and recognition techniques. *Summit on Arabic and Chinese Handwriting Recognition*, pages 22–35.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mahboobeh Sadat Kobari, Nima Karimi, Benyamin Pourhosseini, and Ramin Mousa. 2023. weighted capsulenet networks for persian multi-domain sentiment analysis. *arXiv preprint arXiv:2306.17068*.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection–techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.

Mostafa Masumi, Seyed Soroush Majd, Mehrnoush Shamsfard, and Hamid Beigy. 2025. FaBERT: Pre-training BERT on Persian blogs. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 85–96, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Hossein Mirzaee, Javad Peymanfard, Hamid Habibzadeh Moshtaghin, and Hossein Zeinali. 2022. Armanemo: A persian dataset for text-based emotion detection. *arXiv preprint arXiv:2207.11808*.

Ashritha R Murthy and KM Anil Kumar. 2021. A review of different approaches for detecting emotion from text. In *IOP Conference Series: Materials Science and Engineering*, volume 1110, page 012009. IOP Publishing.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Zeinab Rajabi and MohammadReza Valavi. 2021. A survey on sentiment analysis in persian: a comprehensive system perspective covering challenges and advances in resources and methods. *Cognitive Computation*, 13(4):882–902.

Nazanin Sabri, Reyhane Akhavan, and Behnam Bahrak. 2021a. EmoPars: A collection of 30K emotion-annotated Persian social media texts. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 167–173, Online. INCOMA Ltd.

Nazanin Sabri, Ali Edalat, and Behnam Bahrak. 2021b. Sentiment analysis of persian-english code-mixed texts. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–4. IEEE.

Seyedeh S Sadeghi, Hasan Khotanlou, and M Rasekh Mahand. 2021. Automatic persian text emotion detection using cognitive linguistic and deep learning. *Journal of AI and Data Mining*, 9(2):169–179.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.

Taha Shangipour ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. 2022. Pars-ABSA: a manually annotated aspect-based sentiment analysis benchmark on Farsi product reviews. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7056–7060, Marseille, France. European Language Resources Association.

Pragya Singh Tomar, Kirti Mathur, and Ugrasen Suman. 2023. Unimodal approaches for emotion recognition: A systematic review. *Cognitive Systems Research*, 77:94–109.

Changhao Song, Yazhou Zhang, Hui Gao, Ben Yao, and Peng Zhang. 2025. Large language models for subjective language understanding: A survey. *arXiv preprint arXiv:2508.07959*.

Jie Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1):1.

Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca Passonneau, and Shomir Wilson. 2023. The sentiment problem: A critical survey towards deconstructing sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763, Singapore. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ali Yazdani and Yasser Shekofteh. 2022. A persian asr-based ser: modification of sharif emotional speech database and investigation of persian text corpora. *arXiv preprint arXiv:2211.09956*.

Ali Yazdani, Hossein Simchi, and Yasser Shekofteh. 2021. Emotion recognition in persian speech using deep neural networks. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 374–378.

A Survey Analysis

B Prompt Templates

B.1 Prompt Templates for EA

- **Template (I):** Given a text, identify the main emotion expressed. You have to pick one of the following seven emotions: sadness, hate, anger, happiness, fear, surprise, or other. Only answer with emotion and omit explanations. Emotion:

- **Template (II):** You will be presented with a given comment sourced from X, Instagram, or Digikala. Pick one emotion from sadness, hate, anger, happiness, fear, surprise, or other that describes the emotion of the tweet or comment the best. Your response should only contain one of the emotions. No other output is allowed.

B.2 Prompt Templates for SA

- **Template (I):** Given a text, identify the sentiment expressed. You have to pick one of the following three sentiments: positive, negative, neutral. Only answer with the sentiment and omit explanations. Sentiment:

- **Template (II)** You will be presented with a comment from Digikala. Pick one sentiment

from positive, negative, or neutral that describes the sentiment of the comment the best. Your response should only contain one of sentiment. No other output is allowed.

B.3 Model hyperparameters

B.4 Models

Llama3 (Grattafiori et al., 2024) is an open-access collection of pre-trained and fine-tuned LLMs ranging in scale from 8 billion to 70 billion parameters and launched in September 2024. We examine Llama3-8B model. We use Qwen2-7B-Instruct model that published in November 2024 (Yang et al., 2024). Mistral-7b is also an open-source LM launched in September 2023 (Jiang et al., 2024). Among the models released by Mistral, we test Mixtral-8x7B-Instruct-v0.1, and we access these models via HuggingFace (Wolf et al., 2019).

All responses were collected during July 2024 to March 2025. We run all our experiments on a server with three NVIDIA RTX A6000 and 48GB of RAM.

XLM-RoBERTa The hyperparameters for the XLM-RoBERTa is three epochs, batch size of 16, learning_rate of 2e-5, optimizer of Adam and the maximum length of 128.

B.5 Emotion Analysis

Table 5 shows the performance of the LLMs across different emotions for each dataset.

B.6 Sentiment Analysis

Table 6 shows the performance of the LLMs across different sentiments for each dataset. Mixtral-7B and Llama3-8B can not capture “very negative” and “very positive” labels.

B.7 Toxicity Detection

Table 7 shows the performance of the LLMs across each dataset for detecting offensive/hate speech languagee.

Dataset	Emotion	Mixtral-7B	Llama3-8B	Qwen2-7B	Avg.
Letherlearn	Anger	0.241	0.493	0.358	0.364
	Disgust	0.189	0.056	0.183	0.143
	Fear	0.488	0.461	0.458	0.469
	Happiness	0.423	0.545	0.560	0.509
	Sadness	0.447	0.503	0.511	0.487
	Surprise	0.420	0.264	0.268	0.317
	Other	0.228	0.080	0.345	0.218
Arman	Anger	0.021	0.455	0.456	0.311
	Fear	0.761	0.760	0.733	0.751
	Hate	0.109	0.241	0.441	0.264
	Happiness	0.075	0.521	0.584	0.393
	Sadness	0.414	0.489	0.480	0.461
	Surprise	0.465	0.440	0.483	0.463
	Other	0.231	0.075	0.393	0.233
EmoPars	Anger	0.262	0.307	0.220	0.263
	Fear	0.141	0.162	0.177	0.160
	Hate	0.014	0.046	0.154	0.071
	Happiness	0.247	0.294	0.300	0.280
	Sadness	0.288	0.240	0.256	0.261
	Surprise	0.173	0.066	0.202	0.147

Table 5: F1 Scores for Emotion Analysis Across Datasets and Models with Average.

Dataset	Sentiment	Mixtral-7B	Llama3-8B	Qwen2-7B	Avg.
MirasOpinion	Negative	0.619	0.631	0.656	0.635
	Neutral	0.138	0.498	0.592	0.409
	Positive	0.736	0.812	0.800	0.783
Pars-ABSA	Negative	0.619	0.627	0.616	0.621
	Neutral	0.138	0.332	0.336	0.269
	Positive	0.736	0.741	0.734	0.737
Sentipers	Very Negative	0.000	0.000	0.058	0.019
	Negative	0.560	0.563	0.570	0.564
	Neutral	0.675	0.593	0.664	0.644
	Positive	0.520	0.576	0.586	0.561
	Very Positive	0.000	0.620	0.341	0.320

Table 6: F1 Scores for Sentiment Analysis Across Datasets and Models with Average.

Dataset	Labels	Mixtral-7B	Llama3-8B	Qwen2-7B	Avg.
Pars_OFF	not-offensive	0.841	0.736	0.993	0.857
	offensive	0.640	0.656	0.857	0.718
Phate	not-hate	0.692	0.553	0.720	0.655
	hate	0.673	0.778	0.409	0.620
PHICAD	not-hate	0.911	0.887	0.990	0.929
	hate	0.667	0.653	0.894	0.738

Table 7: Toxicity Detection F1 Scores Across Datasets and Models.