APARSIN: A Multi-Variety Sentiment and Translation Benchmark for Iranic Languages

Sadegh Jafari[1], Tara Azin[2], Farhad Roodi[3,*], Zahra Dehghani Tafti[4,*], Mehrdad Ghadrdan[5,*], Elham Vatankhahan Esfahani[6,*], Aylin Naebzadeh[16,*], Mohammadhadi Shahhosseini[7,*], Ghafoor Khan[12,*], Kazem Forghani[9,*], Danial Namazi[4,*], Seyed Mohammad Hossein Hashemi[8,*], Farhan Farsi[10,*], Mohammad Osoolian[9,*], Maede Mohammadi[11,*], Mohammad Erfan Zare[4,*], Muhammad Hasnain Khan[9,*], Muhammad Hussain[13,*], Nooreen Zaki[14,*], Joma Mohammadi[10,*], Shayan Bali[13,*], Mohammad Javad Ranjbar[4,*], Els Lefever[1,+], Veronique Hoste[1,+]

[1]Ghent University, Belgium, [2]Carleton University, Canada, [3]Purdue University, USA, [4]University of Tehran, Iran, [5]Islamic Azad University Science and Research, Iran, [6]Tarbiat Modares University, Iran, [7]University of Milan, Italy, [8]Shahid Beheshti University, Iran, [9]Iran University of Science and Technology, Iran, [10]Amirkabir university of technology, Iran, [11]Ferdowsi university of Mashhad, Iran, [12]Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, [13]Yuan Ze University, Taiwan, [14]Polytechnical University, Xian, China [15]King's College London, England, [16]University of Hull, England

*** Equal contribution. + Supervisor. Correspondence:** sadegh.jafari@ugent.be

Abstract

The Iranic language family includes many underrepresented languages and dialects that remain largely unexplored in modern NLP research. We introduce APARSIN, a multi-variety benchmark covering 14 Iranic languages, dialects, and accents, designed for sentiment analysis and machine translation. The dataset includes both high- and low-resource varieties, several of which are endangered, capturing linguistic variation across them. We evaluate a set of instruction-tuned Large Language Models (LLMs) on these tasks and analyze their performance across the varieties. Our results highlight substantial performance gaps between standard Persian and other Iranic languages and dialects, demonstrating the need for more inclusive multilingual and dialectally diverse NLP benchmarks.

1 Introduction

The Iranian plateau and its surrounding regions are among the most linguistically diverse areas in the world, home to numerous Iranic languages across multiple branches of the Indo-Iranian family (Kent, 1953; Windfuhr, 2009). This linguistic richness has emerged through centuries of cultural exchange, migration, and sustained language contact across contemporary Iran, Afghanistan, Pakistan, and parts of Iraq (Witzel, 2003; Windfuhr, 2006). While widely spoken languages such as Iranian Persian, Dari, and Pashto serve as lingua francas and administrative languages across much of the region, dozens of smaller language varieties and dialects remain vital to the cultural identity of millions of speakers. Despite this linguistic diversity and extensive prior work on Standard Persian (Farsi et al., 2025a; Abaskohi et al., 2024), Iranic languages other than Standard Persian have been severely underrepresented in natural language processing (NLP) research (Kamaly, 2025), especially in interpretive tasks such as translation and sentiment analysis.

Sentiment analysis is a core area of NLP research with applications across e-commerce, social media monitoring, political analysis, and digital humanities (Liu, 2015). The widespread use of social media platforms has generated large amounts of opinionated text across languages worldwide, creating opportunities to study public sentiment and language use in authentic communicative contexts (Liu, 2022). However, despite substantial advances in sentiment analysis, low-resource and underrepresented languages continue to face significant barriers, mainly due to the scarcity of annotated datasets and the lack of standardized orthographies for wide varieties (Joshi et al., 2020).

For Iranic languages, these challenges are particularly pronounced. Wide varieties lack stan-

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 83–97

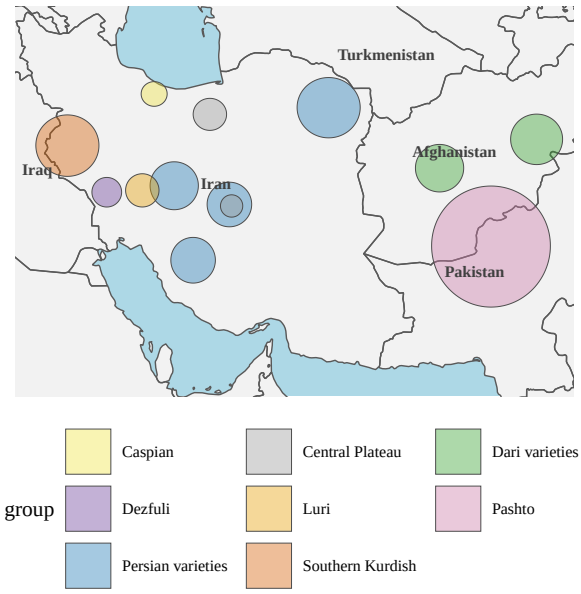

Figure 1: Approximate geographic distribution of Iranic varieties in APARSIN. The map conveys comparative geographic spread rather than exact boundaries, and speaker densities and colors indicate linguistic groups. Based on Glottolog reference points (Hammarström et al., 2025) and generated using Glottospace (Norder et al., 2022).

dardized writing systems (Farsi et al., 2025b) and have historically been transmitted through oral tradition. When written, they typically rely on adapted forms of the Perso-Arabic script with varying orthographic conventions, which often result in inconsistency in digital text. Today, social media serves as an important platform for written communication in these varieties and dialects, yet speakers frequently code-switch between standard Persian and local forms, adapt Persian orthography to represent dialectal features, or use non-standard spellings. These characteristics make sentiment analysis for Iranic varieties particularly challenging. To the best of our knowledge, no comprehensive benchmark currently exists to support research in this domain. You can find our GitHub page here.[1]

To address this gap, we present APARSIN [2], the first multi-variety sentiment and translation analysis benchmark for Iranic languages. The dataset comprises 1,400 annotated social media comments across 14 Iranic languages and dialects: Pashto, Hazaragi, Kabuli Dari, Standard Persian, Shirazi, Khorasani, Esfahani, Yazdi, Kalhori Kurdish, Luri Bakhtiari, Dezfuli, Tonekaboni, Semnani, and Zoroastrian Dari. The dataset represents both major branches of the Iranian language family (Western and Eastern Iranic) and includes both widely spoken languages with millions of speakers (Persian, Pashto, Kurdish) and endangered dialects and language varieties that have received no computational attention (Semnani, Tonekaboni, Zoroastrian Dari). Each sample is annotated for sentiment polarity (positive, negative, neutral) by native speakers and includes translations into Persian and English as well as transliterations for cross-variety and cross-lingual analysis. Figure 1 illustrates the geographic distribution of the languages included in our dataset.

Our contributions are as follows:

- We introduce APARSIN, the first sentiment analysis benchmark covering 14 Iranic languages and dialects, consisting of 1,400 annotated social media comments with corresponding translations.
- Our work presents comprehensive baseline experiments on language identification, machine translation, and sentiment classification using instruction-tuned LLMs for these underrepresented dialects and language varieties.
- The dataset, annotation guidelines, and evaluation code will be released publicly to support future research on sentiment analysis and other NLP tasks for low-resource languages.

2 Related Work

While sentiment analysis achieves high performance for major languages, low-resource and underrepresented languages face substantial challenges due to noisy social media text, non-standardized orthography, and limited training data (Joshi et al., 2020). Developing sentiment analysis resources for such languages remains critical for supporting their speaker communities and advancing cross-linguistic understanding of sentiment expression. For Iranic varieties beyond standard Persian, computational resources remain severely limited. The primary focus of existing work has been on standard Persian spoken in Iran, with several sentiment and emotion analysis datasets developed in recent years. Notable examples include ArmanEmo

[1] https://github.com/SilkRoadAparsin

[2] Pronounced *apārsin* اَپارسین. The name is adapted from Pahlavi sources, meaning "above Simorgh's flight range", where Simorgh is a legendary bird in Persian mythology (Bahar, 1995). The name is used descriptively to evoke the highland regions extending across the Hamoun Lake area (in Iran) and the Hindu Kush.

(Mirzaee et al., 2025) and PersianEmo (Hussiny et al., 2024) for emotion analysis, and SentiFars (Dehkharghani, 2019) for sentiment polarity. However, these resources focus exclusively on standard varieties and do not address the substantial linguistic diversity present within the Iranian language family. Beyond standard Persian, sentiment analysis work for other Iranic varieties is extremely sparse. For Pashto, Kamal et al. (2016) developed a lexicon-based system achieving 73.2% accuracy on social media text. For Kurdish, Badawi (2023) introduced the KMD emotional dataset for Sorani (Central Kurdish), but other Kurdish branches, such as Northern and Southern Kurdish, remain unaddressed. For Central Plateau languages (Semnani, Zoroastrian Dari), Luri varieties, and Caspian languages like Mazandarani, no sentiment resources exist whatsoever. Large-scale multilingual sentiment benchmarks provide important precedents for our work. AfriSenti (Muhammad et al., 2023) covers 14 African languages, and MD-ArSenTD (Baly et al., 2017) addresses multiple Arabic dialects, demonstrating both the feasibility and importance of creating sentiment resources for underrepresented language communities. These efforts highlight common challenges such as non-standardized orthographies, limited digital corpora, and the need for native speaker annotation. Computational approaches to sentiment analysis have also evolved alongside these dataset development efforts. Early work was largely based on rule-based and dictionary-based methods (Mohammad et al., 2013; Taboada et al., 2011; Turney, 2002). Subsequent research shifted toward classical machine learning approaches that relied on manually engineered features (Agarwal and Mittal, 2016; Le and Nguyen, 2015). Advances in deep learning (Yadav and Vishwakarma, 2020; Zhang et al., 2018) and the adoption of pretrained language models have since reshaped the field. Current state-of-the-art systems employ multilingual pretrained models such as XLM-R (Conneau et al., 2020) and mDeBERTaV3 (He et al., 2021), as well as instruction-tuned LLMs that demonstrate strong performance across diverse languages and domains (Zhang et al., 2024). In this work, we evaluate several instruction-tuned LLMs on sentiment classification, language identification, and machine translation tasks for Iranic varieties, providing the first comprehensive assessment of these models' capabilities for this language family.

3 Dataset Overview

APARSIN includes 14 Iranic languages and dialects that represent both major branches of the Iranian language family: Western and Eastern Iranic. The selected varieties cover a wide geographic range, from the mountainous regions of Afghanistan and Pakistan to Iran's verdant Caspian coast, with representation extending into Iraq. This selection includes both major languages with millions of speakers (e.g., Persian, Pashto, Kurdish) and smaller, often endangered varieties (e.g., Semnani, Tonekaboni, Zoroastrian Dari) that have received little to no attention in NLP research. Despite shared Iranic roots, these languages show substantial diversity, which makes them well-suited for studying sentiment analysis across related yet distinct varieties. Table 1 provides an overview of all languages included in the dataset, their classifications, and geographic distributions. Representative examples for each language variety are provided in Table 8 in Appendix.

3.1 Language Family

The Iranian language family, which forms the western branch of the Indo-Iranian languages within Indo-European, is represented in our dataset through three Western Iranic subgroups and one Eastern Iranic branch:

Central Plateau Iranic languages (Semnani and Zoroastrian Dari) belong to the Northwestern Iranic branch and are spoken in central Iran. These languages are associated with some of the longest-standing communities of Iranic speech in the region (see Appendix A). They preserve historically archaic features and have received little to no attention in computational research, in part due to their smaller speaker populations.

Southwestern Iranic constitutes the largest group in our dataset, including Persian and its regional varieties (Kabuli, Hazaragi, Shirazi, and Khorasani), Luri, Dezfuli, and Kalhori (a variety of Southern Kurdish within the Kurdic branch). Standard Persian (also known as Iranian Persian) serves as the standard literary language and lingua franca across much of the region.

Western Iranic (Caspian) is represented by Tonekaboni, a Caspian variety spoken in Mazandaran Province that is commonly treated as distinct from Mazandarani dialects.

Eastern Iranic is represented by Pashto, one of the major languages of Afghanistan and Pakistan,

Language / Variety	ISO	Iranic branch	Region(s)
Persian (Iranian variety)	pes	West Iranic - Southwestern	Iran
Shirazi	pes	West Iranic - Southwestern	Southern Iran (Fars)
Yazdi	pes	West Iranic - Southwestern	Central Iran (Yazd)
Esfahani	pes	West Iranic - Southwestern	Isfahan
Khorasani	pes	West Iranic - Southwestern	Northeastern Iran (Greater Khorasan)
Kabuli	prs	West Iranic - Southwestern	Afghanistan (Kabul and surrounding areas)
Hazaragi	haz	West Iranic - Southwestern	Afghanistan (central highlands)
Pashto	pus	East Iranic	Afghanistan, Pakistan
Kalhori	sdh	West Iranic - Kurdic	Western Iran (Kermanshah), Iraq (Khanaqin)
Luri Bakhtiari	bqi	West Iranic - Southwestern	Southwestern Iran (Zagros region)
Dezfuli	def	West Iranic - Southwestern	Southwestern Iran (Khuzestan)
Tonekaboni	mzn	West Iranic - Caspian	Northern Iran (Caspian coast, Mazandaran)
Semnani	smy	West Iranic - Central Plateau	Central Iran (Semnan)
Zoroastrian Dari	gbz	West Iranic - Central Plateau	Iran (Yazd, Kerman)

Table 1: Iranic languages and dialects included in our dataset. ISO 639-3 codes are reported at the language level (varieties without distinct ISO codes share the parent language code).

with substantial phonological and morphological differences from the Western Iranic varieties.

3.2 Geographic Scope

Geographically, the dataset covers Iran, Afghanistan, Iraq, and Pakistan. Many of these languages show significant dialectal variation across regions. Persian varieties, for instance, differ substantially between Iran (Shirazi, Khorasani, Esfahani, Standard Persian, Yazdi) and Afghanistan (Kabuli, Hazaragi), showing both geographic separation and distinct ethnic community identities. Kurdish varieties are also found across multiple countries, with Kalhori (Southern Kurdish) spoken primarily in Iran's Kermanshah province and extending into Iraq's Khanaqin region (Azin and Ahmadi, 2021).

3.3 Writing Systems

All languages in our dataset use the Perso-Arabic script, except Zoroastrian Dari, which is provided only using transliteration due to its exclusively oral transmission among the community from which our data were collected. It includes additional letters not found in Arabic to represent phonemes such as /p/, /tʃ/, /ʒ/, and /g/. While the script is shared across varieties, individual languages use distinct orthographic conventions and, in some cases, additional diacritics reflecting local phonological features. The widespread use of Perso-Arabic script shows shared historical and cultural influences, as well as Persian's longstanding role as a prestige and administrative language. Moreover, since many smaller Iranic languages and dialects have historically been transmitted through oral traditions, in contemporary social media contexts, speakers adapt standard Persian orthography to represent such varieties. This results in non-standard spellings, orthographic variation, and code switching between standard and dialectal forms.

4 Dataset Collection and Processing

Collecting authentic and representative data for Iranic languages poses significant challenges, particularly for low-resource and endangered varieties. Many of these languages have a limited digital presence, lack standardized orthography, and are primarily transmitted through oral traditions. To address these challenges and ensure ecological validity, we adopted a multi-step data collection strategy that combines existing resources, automated web crawling, and manual data collection with the help of native speakers.

4.1 Speaker Recruitment

Native speakers of different Iranic language varieties were recruited through social media platforms such as LinkedIn and other community-based online networks. All recruited speakers participated voluntarily. As a form of acknowledgment and compensation, contributors who provided substantial data are included as co-authors in this paper.

This recruitment strategy was chosen to ensure the collection of real-world, authentic language use as it naturally appears in social media and everyday communication. Such an approach is particularly important for low-resource and endangered Iranic languages, for which curated datasets are scarce or entirely unavailable, and where linguistic knowledge is often maintained within small speaker communities.

4.2 Dataset Collection

We employed three complementary data collection approaches depending on the availability and digital footprint of each language variety.

4.2.1 Published Datasets

For well-resourced and standardized Iranic languages, such as Standard Persian and Pashto, we rely on existing publicly available datasets. These resources provide relatively large volumes of high-quality textual data and serve as a solid foundation for languages with established writing systems and sufficient online presence. In this work, we make use of the English–Pashto Language Dataset (EPLD) (Khan et al., 2025), the Pashto–English Bilingual Sentiment Corpus[3], as well as two large-scale Persian sentiment analysis datasets, namely Digikala Sentiment [4] and SnappFood Sentiment (Farahani et al., 2020).

4.2.2 Web Crawling

For under-resourced languages, dialects, and accent varieties, we developed an automated web-crawling framework to gather naturally occurring language data from the internet. The framework targets sources such as social media platforms, personal and community blogs, as well as linguistic, cultural, and regional websites.

Automated Crawling Pipeline. The crawling process followed a multi-stage pipeline powered by LLMs. First, we prompted an LLM to generate language-specific search keywords, including alternative spellings, colloquial forms, and community-specific identifiers. These keywords were then used to query search engines and online platforms automatically. Next, for each retrieved document, the LLM was used to extract candidate sentences written in the target language variety, along with their corresponding translations when available. All experiments reported in this work used GPT-5 (OpenAI, 2025a) as the underlying LLM.

Human Annotation. To ensure data quality, we employed native speakers of each language variety to annotate the collected samples manually. For each sentence, annotators verified (i) whether the original utterance is valid and genuinely belongs to the target language variety, and (ii) whether the provided translation is correct. Both checks were annotated using binary True/False labels. Table 2 reports the aggregated results of this annotation process.

[3] https://www.kaggle.com/datasets/farhadkhan66/pashto-translated-corpus

[4] https://www.digikala.com/opendata/

Results and Discussion. Table 2 summarizes the crawling and annotation outcomes for different language varieties. The number of collected samples varies substantially across languages, reflecting differences in online presence and community activity. Several varieties, such as *Dezfuli*, *Khorasani*, and *Yazdi* exhibit very high original validity rates, indicating that the LLM-guided keyword generation and sentence extraction were effective in identifying genuine language samples.

In contrast, languages such as *Luri* and *Semnani* show lower validity and translation correctness rates, highlighting the challenges of noisy web data and limited standardized written forms. Notably, for *Kalhori*, while most original utterances were valid, translation correctness was considerably lower, suggesting difficulties in obtaining reliable translations for certain varieties even when the original text is available. Finally, for some endangered languages with extremely limited or nonexistent online presence (e.g., Zoroastrian varieties), web crawling yielded little to no usable data, underscoring the limitations of purely web-based collection methods for severely under-documented languages.

Language	Samples	Orig. Valid	Trans. Correct
Dezfuli	764	99.738%	85.602%
Hazaragi	383	80.940%	58.486%
Esfahani	294	83.333%	82.993%
Khorasani	1216	99.753%	99.342%
Luri	1040	31.538%	31.635%
Semnani	470	60.213%	58.511%
Shirazi	563	65.187%	71.048%
Kalhori	403	99.504%	25.310%
Tonekaboni	437	79.863%	64.073%
Yazdi	697	98.278%	96.987%

Table 2: Dataset statistics per language variety. **Orig. Valid** denotes the proportion of samples whose original utterance was verified by native speakers as valid and belonging to the target language, while **Trans. Correct** indicates the proportion of samples with a correct human-verified translation.

4.2.3 Manual Data Collection

For languages where automated methods were ineffective, we relied on manual data collection by volunteer native speakers. Contributors were asked to collect short, common, and naturally occurring text samples from social media platforms such as Facebook, Twitter/X, and messaging forums, or to provide original examples representative of everyday usage. Clear guidelines and collection protocols were provided to ensure consistency, ethical data

handling, and linguistic authenticity. This manual approach was essential for preserving endangered languages that are rarely written and poorly represented in digital spaces.

4.3 Topic Modeling and Sample Selection

We apply a topic modeling pipeline based on BERTopic (Grootendorst, 2022) to discover semantic patterns across multilingual Iranic language data. Sentence-level representations are obtained using the BGE-M3 (Chen et al., 2024) multilingual embedding model, while topic representations are generated using GPT-4o-Mini (OpenAI, 2025b) to improve interpretability. Dimensionality reduction is performed using UMAP (McInnes et al., 2018) with 15 nearest neighbors, a cosine distance metric, and a 10-dimensional projection for clustering, followed by a two-dimensional projection for visualization. Topics are identified using HDBSCAN (Campello et al., 2013) with a minimum cluster size of 10, enabling density-based discovery without predefining cluster shapes. The model is configured to extract 10 distinct topics, corresponding to the 10 languages and dialects in the dataset. For balanced sample selection, one representative sample is selected for each topic–language pair, resulting in a total of 100 selected samples (for more details see Appendix B).

4.4 Translation and Transliteration Approaches

In order to create a comprehensive and high-quality dataset for low-resource Iranic varieties, we employed a combination of translation and transliteration strategies to capture authentic, real-world language use from social media.

4.4.1 Translation Approach

To address the limited digital presence of low-resource and endangered Iranic varieties, part of the dataset was created through translation. For varieties lacking sufficient naturally occurring written data, sentences were translated from high-resource languages such as Persian and English. Translations were performed by native speakers and professional translators, depending on availability and linguistic needs. The translation direction was chosen on a case-by-case basis to ensure semantic accuracy and cultural relevance. Quality was ensured through cross-checking by additional native speakers, targeted reviews by linguistically trained annotators, and consistency checks across sentiment annotations.

4.4.2 Transliteration Approach

All sentences were also transliterated into a unified Latin-based format using a standardized scheme adapted from the Iranian Studies transliteration guidelines[5]. Annotators followed the guidelines to maintain consistency across languages.

This transliteration supports cross-linguistic analysis, improves accessibility for researchers unfamiliar with Perso-Arabic scripts, facilitates integration with NLP models, and enables future reuse of the dataset in multilingual settings.

4.5 Sentiment Annotation Process

After selecting 100 samples and translating them into each target language, we asked three native volunteer annotators per language to label the sentiment of each sentence as *Negative*, *Neutral*, or *Positive*. To assess annotation reliability, we computed inter-annotator agreement using Krippendorff's α (KA) and the average pairwise Cohen's κ (CK).

Table 3 reports the agreement scores across the Iranic language varieties. We observe substantial agreement for Tonekaboni ($\alpha = 0.913$), indicating highly consistent sentiment interpretation among annotators. Moderate agreement is achieved for languages such as Kalhori, Esfahani, and Semnani, while lower agreement is observed for Pashto, Shirazi, and Dezfuli. These variations likely reflect differences in dialectal ambiguity, sentiment expression, and the limited availability of standardized sentiment cues in low-resource language varieties. Overall, the results highlight both the feasibility of sentiment annotation and the inherent challenges of achieving high agreement across diverse Iranic languages.

5 Experiments

We conduct our experiments using eight instruction-tuned LLMs spanning four model families: OpenAI: GPT-4o and GPT-4o-mini (Achiam et al., 2023); Google: Gemma-3-12B-IT and Gemma-3-27B-IT (Team et al., 2025); Meta: Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Grattafiori et al., 2024); and Qwen: Qwen3-14B and Qwen3-32B (Yang et al., 2025).

[5] https://github.com/SilkRoadAparsin/Translation

Language	KA	CK
Tonekaboni	0.913	0.913
Zoroastrian Dari	0.423	0.444
Kalhori	0.579	0.581
Pashto	0.197	0.215
Esfahani	0.541	0.544
Shirazi	0.384	0.391
Semnani	0.510	0.510
Dezfuli	0.330	0.332

Table 3: Inter-Annotator Agreement (IAA) scores for sentiment annotation across Iranic language varieties. **KA** denotes Krippendorff's α, measuring overall annotation reliability across multiple annotators, while **CK** denotes the average pairwise Cohen's κ, reflecting annotator consistency at the pair level. Higher values indicate stronger agreement.

5.1 Language Detection

We evaluate LLMs on identifying the *language* (L), *dialect* (D), and *accent* (A) of samples in APARSIN, a challenging setting due to the close relatedness of Iranic varieties, shared Perso-Arabic script, and non-standardized orthography. As shown in Table 4, all models perform poorly at the language level (macro-F1 ≤ 0.18), frequently confusing closely related dialects such as Persian, Kurdish, and Luri. Dialect identification shows slightly higher but still limited performance (macro-F1 up to 0.22), with better results for more lexically distinctive varieties such as Semnani and Dezfuli, while Persian regional dialects are often conflated. Accent identification achieves the highest scores (macro-F1 up to 0.48), particularly for varieties with salient non-standard lexical or phonological cues reflected in writing (e.g., Hazaragi, Southern Kurdish), whereas *General* accents remain difficult to detect.

5.2 Machine Translation

Each model was prompted to translate sentences from the Iranic languages into English, and the resulting machine-generated translations were compared against the gold English references available in our dataset. Translation quality was measured using BERTScore by computing the semantic similarity between the embeddings of the reference English translations and the model-generated English outputs, as well as BLEU for surface-level n-gram overlap. As shown in Tables 5 and 6, GPT-4o and LLaMA-70B consistently achieve the highest translation quality across Iranic languages, while Qwen models perform poorly. Languages such as Khorasani and Pashto yield better scores under both BERTScore and BLEU.

5.3 Sentiment Classification

As shown in Table 7, larger models generally achieve higher macro F1 scores across languages, dialects, and accents, with GPT-4o and Gemma-27B obtaining the best average performance, while smaller and Qwen models show noticeably lower results.

6 Conclusion

In this work, we presented APARSIN, a multi-variety benchmark for sentiment analysis and machine translation across 14 Iranic languages and dialects. By explicitly modeling variation at the language, dialect, and accent levels, our dataset exposes significant disparities in model performance, particularly for low-resource and endangered varieties. Experimental results with instruction-tuned LLMs show that strong performance on standard Persian does not generalize to other Iranic varieties. We hope that APARSIN will encourage future research on inclusive evaluation, data collection, and modeling approaches for underrepresented language communities.

Limitations

This work has several limitations. First, due to recent internet blackouts in Iran, communication with a subset of annotators was temporarily disrupted, resulting in the loss of a small portion of annotation results. These missing annotations will be recovered and incorporated into a future release of the dataset once connectivity with Iran is fully restored. Second, although APARSIN covers a broad range of Iranic languages and dialects, the dataset remains relatively small in scale, with 1,400 annotated samples. This limits the extent to which conclusions can be generalized, particularly for training data-intensive models. Finally, our experiments focus on sentiment analysis and machine translation using instruction-tuned LLMs. While these models provide a strong baseline, they may not reflect the performance of task-specific or fine-tuned models, nor do they fully capture other important NLP challenges for Iranic languages, such as morphological analysis or syntactic variation. Future work should address these limitations by expanding the dataset and incorporating additional tasks.

Lang	Dialect	Accent	Models																							
			GPT-4o			GPT-4o-mini			Qwen-32B			Qwen-14B			Gemma-27B			Gemma-12B			LLaMA-70B			LLaMA-8B		
			L	D	A	L	D	A	L	D	A	L	D	A	L	D	A	L	D	A	L	D	A	L	D	A
Caspian	Mazandarani	Tonekaboni	0.00	0.29	0.09	0.00	0.15	0.30	0.00	0.21	0.23	0.00	0.01	0.00	0.00	0.01	0.28	0.00	0.04	0.20	0.00	0.16	0.32	0.00	0.44	1.00
Central Iran	Zoroastrian Dari	Yazdi	0.00	0.00	1.00	0.00	0.00	0.46	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.33	0.00	0.00	1.00	0.00	0.00	0.49
Kurdish	Southern	Southern	0.32	0.03	1.00	0.24	0.03	1.00	0.14	0.00	0.35	0.01	0.00	0.01	0.49	0.21	1.00	0.17	0.14	1.00	0.24	0.04	1.00	0.23	0.00	1.00
Luri	Southern	General	0.11	0.30	0.00	0.00	0.49	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.33	0.00	0.00	1.00	0.00	0.05	0.25	0.00	0.00	1.00	0.00
Pashto	Central	General	0.33	0.03	0.00	0.23	0.00	0.00	0.15	0.19	0.00	0.02	0.01	0.00	0.50	0.02	0.00	0.19	0.02	0.00	0.24	0.06	0.44	0.16	0.28	0.48
Persian	Dari	General	0.16	0.10	0.00	0.32	0.07	0.00	0.32	0.05	0.19	0.08	0.00	0.00	0.24	0.24	0.19	0.24	0.13	0.23	0.16	0.03	0.00	0.24	0.00	0.03
Persian	Hazaragi	Hazaragi	0.06	0.09	1.00	0.14	0.06	0.49	0.15	0.00	0.37	0.03	0.00	0.07	0.22	0.00	1.00	0.13	0.03	1.00	0.10	0.05	1.00	0.31	0.00	1.00
Persian	Iranian	General	0.50	0.32	0.50	0.25	0.24	0.00	0.36	0.15	0.13	0.07	0.03	0.03	0.33	0.29	0.19	0.50	0.26	0.30	0.33	0.33	0.33	0.33	0.50	0.19
Persian	Iranian	Isfahani	0.14	0.10	0.00	0.15	0.09	0.12	0.13	0.08	0.00	0.03	0.03	0.00	0.32	0.04	0.00	0.19	0.02	0.01	0.22	0.14	0.02	0.32	0.24	0.00
Persian	Iranian	Shirazi	0.19	0.14	0.09	0.24	0.11	0.00	0.18	0.17	0.03	0.04	0.00	0.00	0.24	0.07	0.07	0.24	0.06	0.06	0.22	0.15	0.01	0.24	0.24	0.01
Persian	Iranian	Yazdi	0.16	0.17	0.00	0.22	0.11	0.00	0.17	0.17	0.00	0.01	0.01	0.00	0.23	0.10	0.00	0.23	0.05	0.00	0.18	0.20	0.00	0.22	0.31	0.00
Semnani	Semnani	Semnani	0.00	0.48	1.00	0.00	0.49	1.00	0.00	0.35	0.38	0.00	0.07	0.04	0.00	1.00	1.00	0.00	0.49	1.00	0.00	0.23	1.00	0.00	0.50	1.00
Unclassified	Dezfuli	Dezfuli	0.00	0.46	1.00	0.01	0.47	0.50	0.00	0.21	0.33	0.00	0.00	0.01	0.00	0.43	1.00	0.00	0.41	1.00	0.00	0.50	1.00	0.00	0.18	1.00
	Macro-F1		0.14	0.22	0.42	0.18	0.20	0.48	0.13	0.14	0.31	0.02	0.02	0.02	0.15	0.18	0.45	0.13	0.17	0.47	0.15	0.19	0.45	0.16	0.22	0.47

Table 4: Evaluation of LLMs on the APARSIN dataset for identifying *language* (L), *dialect* (D), and *accent* (A) across Iranic varieties.

Language	GPT-4o	GPT-4o-mini	Qwen-32B	Qwen-14B	Gemma-27B	Gemma-12B	LLaMA-70B	LLaMA-8B
Shirazi	0.4590	0.4489	-0.2614	-0.2790	0.0906	0.4514	0.3887	0.1631
Yazdi	0.4660	0.4147	-0.2208	-0.2483	0.0554	0.4572	0.3463	0.0894
Esfahani	0.4339	0.4192	-0.2661	-0.2890	0.0372	0.4381	0.3360	0.0784
Khorasani	0.6259	0.6364	-0.2284	-0.2363	0.4200	0.6870	0.6536	0.4685
Kabuli	0.3014	0.3208	-0.2694	-0.2805	0.1508	0.3451	0.3254	0.1877
Hazaragi	0.3947	0.3482	-0.2412	-0.2747	0.0496	0.3732	0.3414	0.0258
Pashto	0.5747	0.5540	-0.2327	-0.2937	0.4033	0.5652	0.5212	0.3158
Kalhori	0.3168	0.2493	-0.3115	-0.3203	0.1073	0.2672	0.2332	0.0367
Luri Bakhtiari	0.3252	0.2720	-0.2469	-0.2687	0.0597	0.3011	0.2567	0.1014
Dezfuli	0.3695	0.3221	-0.2747	-0.2735	-0.0260	0.3520	0.2723	0.0280
Tonekaboni	0.4452	0.3758	-0.2855	-0.2924	-0.0095	0.4281	0.3602	0.0109
Semnani	0.3298	0.2688	-0.3080	-0.3146	-0.0823	0.2764	0.2495	-0.0264
Zoroastrian Dari	0.2306	0.1870	-0.2145	-0.2276	-0.1515	0.2276	0.1718	-0.0036

Table 5: BERT Score results for translation from Iranic languages into English across different LLMs.

Language	GPT-4o	GPT-4o-mini	Qwen-32B	Qwen-14B	Gemma-27B	Gemma-12B	LLaMA-70B	LLaMA-8B
Shirazi	0.2078	0.1899	0.0016	0.0029	0.0258	0.1091	0.0950	0.0211
Yazdi	0.2073	0.1556	0.0028	0.0029	0.0187	0.1005	0.0826	0.0216
Esfahani	0.1482	0.1602	0.0016	0.0019	0.0177	0.0942	0.0714	0.0219
Khorasani	0.3642	0.4195	0.0101	0.0077	0.1368	0.3425	0.2909	0.1326
Kabuli	0.1213	0.1647	0.0028	0.0030	0.0603	0.1170	0.1130	0.0492
Hazaragi	0.1049	0.0783	0.0012	0.0004	0.0147	0.0585	0.0553	0.0061
Pashto	0.2671	0.2457	0.0030	0.0029	0.1173	0.1782	0.1959	0.0613
Kalhori	0.0599	0.0294	0.0008	0.0012	0.0201	0.0313	0.0346	0.0074
Luri Bakhtiari	0.0755	0.0461	0.0010	0.0007	0.0128	0.0505	0.0555	0.0093
Dezfuli	0.1105	0.0734	0.0017	0.0016	0.0153	0.0854	0.0468	0.0132
Tonekaboni	0.1437	0.1029	0.0021	0.0016	0.0130	0.0791	0.0789	0.0088
Semnani	0.0545	0.0608	0.0006	0.0004	0.0058	0.0260	0.0205	0.0022
Zoroastrian Dari	0.0431	0.0152	0.0005	0.0003	0.0024	0.0123	0.0176	0.0025

Table 6: BLEU scores for translation from Iranic languages into English.

Language	GPT-4o	GPT-4o-mini	Qwen-32B	Qwen-14B	Gemma-27B	Gemma-12B	LLaMA-70B	LLaMA-8B
Tonekaboni	0.6363	0.6286	0.2132	0.3033	0.6551	0.6412	0.5044	0.3104
Zoroastrian Dari	0.0550	0.0550	0.2258	0.2782	0.0550	0.0550	0.1925	0.1082
Kalhori	0.4557	0.5305	0.2938	0.2342	0.6188	0.3678	0.4666	0.2507
Pashto	0.6849	0.7105	0.1764	0.1705	0.6576	0.6618	0.4242	0.3489
Esfahani	0.6410	0.6413	0.3160	0.3754	0.5847	0.5676	0.5489	0.3992
Shirazi	0.6693	0.6500	0.3252	0.2796	0.6552	0.6730	0.5381	0.3237
Semnani	0.5199	0.4847	0.2582	0.2772	0.4909	0.4696	0.3692	0.3544
Dezfuli	0.5999	0.5490	0.2828	0.2670	0.5815	0.5820	0.3752	0.3955
AVERAGE	**0.5327**	**0.5312**	**0.2614**	**0.2732**	**0.5374**	**0.5023**	**0.4274**	**0.3114**

Table 7: Sentiment performance across Iranic dialects and related languages. All scores are macro F1.

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. *Preprint*, arXiv:2404.02403.

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

Basant Agarwal and Namita Mittal. 2016. Machine learning approach for sentiment analysis. In *Prominent Feature Extraction for Sentiment Analysis*, pages 21–45. Springer.

Zahra Azin and Sina Ahmadi. 2021. Creating an electronic lexicon for the under-resourced southern varieties of kurdish language. In *Proceedings of Seventh Biennial Conference on Electronic Lexicography (eLex 2021)*.

Soran Badawi. 2023. Kmd: A new kurdish multilabel emotional dataset for the kurdish sorani dialect. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing*, pages 308–315, Online. Association for Computational Linguistics.

Mehrdad Bahar. 1995. *Bundahišn*. Toos Publications, Tehran, Iran.

Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

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.

Rahim Dehkharghani. 2019. Sentifars: A persian polarity lexicon for sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–12.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *arXiv*, abs/2005.12515.

Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli Sadraei-Javaheri, Shayan Bali, Amir Hossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, and Saeedeh Momtazi. 2025a. MELAC: Massive evaluation of large language models with alignment of culture in Persian language. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1933–1950, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Farhan Farsi, Parnian Fazel, Farzaneh Goshtasb, Nadia Hajipour, Sadra Sabouri, Ehsaneddin Asgari, and Hossein Sameti. 2025b. PahGen: Generating Ancient Pahlavi text via grammar-guided zero-shot translation. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 171–182, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Mehrdad Ghadrdan. 2007. *A Linguistic and Grammatical Study of the Zoroastrian Dialect of Sharifabad, Ardakan, Yazd*. Rakhshid Publishing, Shiraz, Iran.

Saloumeh Gholami. 2018. Remnants of zoroastrian dari in the colophons and sālmargs of iranian avestan manuscripts. *Iranian Studies*, 51(2):195–211.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2. http://glottolog.org. Accessed 2025-12-31.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *Proceedings of the International Conference on Learning Representations*.

Mohammad Ali Hussiny, Mohammad Arif Payenda, and Lilja Øvrelid. 2024. Persianemo: Enhancing farsidari 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, Italy. ELRA and ICCL.

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.

Uzair Kamal, Imran Siddiqi, Hammad Afzal, and Arif ur Rahman. 2016. Pashto sentiment analysis using lexical features. In *Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, Tebessa, Algeria.

Arta Modarres Kamaly. 2025. Towards inclusive nlp: Evaluating llms on low-resource indo-iranian languages. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*.

Roland G Kent. 1953. *Old Persian: Grammar. Texts. Lexicon*, volume 33. American Oriental Society.

Rabia Khan, Huzaifa Saleem Khan, and Shireen Ijaz. 2025. English-pashto language dataset (epld). *Mendeley Data*, (V1).

Binh Le and Huy Nguyen. 2015. Twitter sentiment analysis using machine learning techniques. In *Advanced Computational Methods for Knowledge Engineering*, volume 358 of *Advances in Intelligent Systems and Computing*, pages 279–288. Springer.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge, UK.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Leland McInnes, John Healy, and James Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction. arxiv. *arXiv preprint arXiv:1802.03426*, 10.

H. Mirzaee, J. Peymanfard, H. Habibzadeh Moshtaghin, and 1 others. 2025. Armanemo: A persian dataset for text-based emotion detection. *Language Resources and Evaluation*, 59:2565–2587.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiao-Dan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Sietze Norder, Laura Becker, Hedvig Skirgård, Leonardo Arias, Alena Witzlack-Makarevich, and Rik van Gijn. 2022. glottospace: R package for the geospatial analysis of linguistic and cultural data. *Journal of Open Source Software*, 7(77):4303.

OpenAI. 2025a. Gpt-5. OpenAI model documentation. https://platform.openai.com/docs/models/gpt-5.

OpenAI. 2025b. Introducing gpt-4o-mini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed: 2025-06-04.

Ebrahim Rezapour and Mona Pishgou. 2013. A typological study of causative constructions in the semnani language. *Journal of Linguistic and Rhetorical Studies*, (23):175–214.

Arash Salar. 2019. The case system in the semnani language within a categorial framework. M.a. thesis, Faculty of Persian Literature and Foreign Languages, University of Tehran, Tehran, Iran.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

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

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Gernot Windfuhr. 2009. *The Iranian Languages*, volume 2009. Routledge London.

Gernot L. Windfuhr. 2006. Iran vii. non-iranian languages (6) in islamic iran. In *Encyclopaedia Iranica*, volume 13, pages 393–396. Encyclopaedia Iranica Foundation.

Michael Witzel. 2003. *Linguistic evidence for cultural exchange in prehistoric western Central Asia*. 129. Department of East Asian Languages and Civilizations, University of ….

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, 53(6):4335–4385.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

A Endangered languages in the Dataset

Zoroastrian Dari (also known as Behdini or Gavruni) is an endangered Iranic language spoken by Zoroastrian communities primarily in Yazd and surrounding areas of Iran, with smaller populations in Kerman and Tehran. The language exhibits substantial dialectal variation and is used almost exclusively in spoken form, with no standardized written tradition (Gholami, 2018). Within the Yazd Province, Zoroastrian Dari includes multiple dialects specific to individual villages and Zoroastrian quarters. These varieties differ in phonetics, phonology, morphology, syntax, and vocabulary. In this study, we focus specifically on the Zoroastrian dialect of Sharif Abad (Ardakan, Yazd), for which the monograph by Ghadrdan (2007) provides the earliest and most comprehensive linguistic description available to date. Owing to the absence of publicly available corpora and the exclusively oral nature of the dialect, this study relies on examples and elicited materials documented in that work as the primary source for dataset construction.

Sentiment annotation for this dialect was challenging due to its limited use among younger speakers. Since many sentences were translated into Sharif Abad's Zoroastrian Dari and written in our transliteration scheme, annotators were unfamiliar with reading them. We addressed this by reading examples aloud during voice-based annotation sessions. Utterances were delivered with neutral prosody to minimize potential annotation bias, and repetitions were used when necessary to support accurate understanding. These methodological adaptations indicate the challenges of working with an endangered, predominantly oral language and should be considered when interpreting the sentiment annotations.

Semnani is another endangered Iranic language in our dataset spoken in and around the city of Semnan in north central Iran. From a genetic and historical perspective, Semnani is rooted in Parthian (also known as Arsacid Pahlavi) and belongs to the Northwestern branch of Iranic languages (Salar, 2019). The language holds a distinctive position within the Iranian linguistic landscape due to its high degree of structural preservation and authenticity, while at the same time facing the risk of extinction (Rezapour and Pishgou, 2013). Semnani comprises several closely related but distinct varieties spoken in Semnan and neighboring areas, including Sorkhe'i, Lasgerdi, Sangsari, and

Aftarabadi.

The language lacks publicly available corpora and is transmitted primarily through oral use, though written forms use the Perso-Arabic script. However, ongoing community-based revitalization efforts have supported continued engagement with the language, including emerging literacy among younger speakers. Accordingly, sentiment annotation was conducted by three literate community members (two aged 26 and one aged 43), demonstrating the viability of written annotation despite the language's endangered status.

B Topic Modeling and Sample Selection

To identify semantically coherent patterns across Iranic languages and dialects, we employ a topic modeling framework that integrates semantic representation learning, non-linear dimensionality reduction, density-based clustering, and large language model–based topic labeling.

First, all textual samples are encoded into a shared semantic space using a multilingual sentence embedding model. This representation captures cross-lingual and intra-dialectal semantic similarity, enabling meaningful comparison between closely related language varieties as well as more distant ones. The use of a transformer-based encoder ensures robustness to lexical variation and supports low-resource dialects.

Given the high dimensionality of the resulting embeddings, a non-linear manifold learning technique is applied to project the representations into a lower-dimensional space. The dimensionality reduction is guided by neighborhood-based hyperparameters that preserve local semantic structure while maintaining global separability between emerging topics. A higher-dimensional projection is used to support stable clustering, while a two-dimensional projection is reserved exclusively for visualization and qualitative analysis.

Topic discovery is performed using a density-based clustering algorithm. This approach does not assume a predefined number of clusters and is well-suited for data with irregular cluster shapes. A minimum cluster size constraint is imposed to ensure that identified topics correspond to semantically meaningful groupings rather than noise. Samples that do not strongly belong to any dense region are treated as outliers, preventing forced topic assignments.

For topic interpretability, each discovered cluster is labeled using a large language model. Instead of relying solely on keyword frequency, the model generates concise, human-readable topic descriptors based on representative samples within each cluster. This strategy significantly improves interpretability, particularly in multilingual and dialectally diverse settings.

The final output is a two-dimensional semantic map in which samples are grouped and labeled according to their inferred topics. Representative points are selected using medoid-based labeling to reduce visual clutter and enhance readability. This visualization supports informed sample selection and qualitative inspection of linguistic variation across the dataset.

The dataset used in this study covers ten Iranic languages and dialects, including multiple varieties of Persian as well as Pashto, Kurdish, Mazandarani, Semnani, Dezfuli, and Luri. This diversity enables the model to capture both cross-language semantic structure and fine-grained dialectal distinctions.

C Additional Prompt Templates

All classification tasks are framed as **closed-set** problems with deterministic decoding (temperature $= 0$).

C.1 Language, Dialect, and Accent Detection

Language-related identification is evaluated using a three-stage prompting pipeline.

System Message

You are a classifier. Given a text and an allowed label list, choose exactly one label from the list. Return a JSON object.

Stage 1: Language Identification

Task: Identify the language of the given text.
Output format (JSON): {"language": "<LABEL>"}
Constraint: The value for "language" must be exactly one item from the allowed list.
Allowed languages (choose ONE): [...]
Text: <TEXT>

Stage 2: Dialect Identification

Task: Identify the dialect of the text (language is given).
Output format (JSON): {"dialect": "<LABEL>"}
Constraint: The value for "dialect" must be exactly one item from the allowed list.
Language: <LANGUAGE>
Allowed dialects (choose ONE): [...]
Text: <TEXT>

Stage 3: Accent Identification

Task: Identify the accent of the text (language and dialect are given).
Output format (JSON): {"accent": "<LABEL>"}
Constraint: The value for "accent" must be exactly one item from the allowed list.
Language: <LANGUAGE>
Dialect: <DIALECT>
Allowed accents (choose ONE): [...]
Text: <TEXT>

C.2 Translation

System Message

You are a professional translator with expertise in Iranic languages. Just return the translation.

Text Translation

Task: Translate the given text into the target language, considering the provided language, dialect, and accent.
Output format: Plain text (translation only, no explanations).
Source Language: <LANGUAGE>
Dialect: <DIALECT>
Accent: <ACCENT>
Target Language: <TARGET_LANGUAGE>
Text: <TEXT>

C.3 Sentiment Classification

Sentiment classification is conducted using metadata-aware prompting.

System Message

You are a classifier. Given a text and an allowed label list, choose exactly one label from the list. Return a JSON object.

Sentiment Classification

Task: Identify the sentiment of the text (language, dialect, and accent are given).
Output format (JSON): {"sentiment": "<LABEL>"}
Constraint: The value for "sentiment" must be exactly one item from the allowed list.
Language: <LANGUAGE>
Dialect: <DIALECT>
Accent: <ACCENT>
Allowed sentiments (choose ONE): [...]
Text: <TEXT>

Language/Dialect	Example	Transliteration	English	Standard Persian
Semnani	خوشتُنَه بَخُو، مَرتُمُنَه تُون کَه.	khoshtona bakhow, martomona town ka	Eat for yourself, dress for others.	برای خودت بخور، برای مردم بپوش.
Dezfuli	سی چ؟	si che	For what?	برای چه؟
Esfahani	حجمی ساندویچش کلا نصف شده‌س.	hajmi sāndevichesh kolan nesf shodees	The size of his sandwich has completely halved!	حجم ساندویچش کلا نصف شده!
Yazdi	موخوم یققم پاره کنم.	mokhom yagham pāra konam	I want to tear my collar (from despair/anger)!	می‌خواهم یقه‌ام را پاره کنم!
Zoroastrian Dari	not available	tow kovari barem za	From which direction has the sun risen?	آفتاب از کدام سمت درآمده است؟
Tonekaboni	اَما هم آنیم.	amā ham ānim	We are coming too.	ما هم می‌آیم.
Kalhori	ای گیان ئەڕا شارگان.	ey gyān arrā shāragamān	Oh, our dear city.	ای جان برای شهرمان.
Luri Bakhtiari	هر چه دیش عاقل بی خوش کلوعه.	har che deish āqel bi khosh kalu'e	Unlike her mother, who was sensible, she seems to be a fool.	برعکس مادرش که عاقل بود، خودش انگار دیوانه است.
Shirazi	خو پَ چرو نمیوی؟	kho pa chero namioy	Then why don't you come?	آخه پس چرا نمی‌آیی؟
Khorasani	یره چقد خیت رِفت.	yare cheqad khit reft	Bro, that was such a waste!	داداش، چقدر ضایع شد.
Kabuli Dari	خدا خر ره دیده که برش شاخ نداده.	Khodā khar ra dida ke barash shākh nadāda	God knew the donkey well, so He didn't give it horns.	خدا خر را دیده که به او شاخ نداده.
Pashto	هوا ناڅاپه سړه شوه۔	hawā nātsāpeh sra shoeh	The weather suddenly became cold.	هوا ناگهان سرد شد.
Hazaragi	کلو جالب خاد بود.	kalo jāleb khād bud	It would be very interesting.	خیلی جالب خواهد بود.

Table 8: Examples across varieties: original sentence (Perso-Arabic script), transliteration, English translation, and Standard Persian.

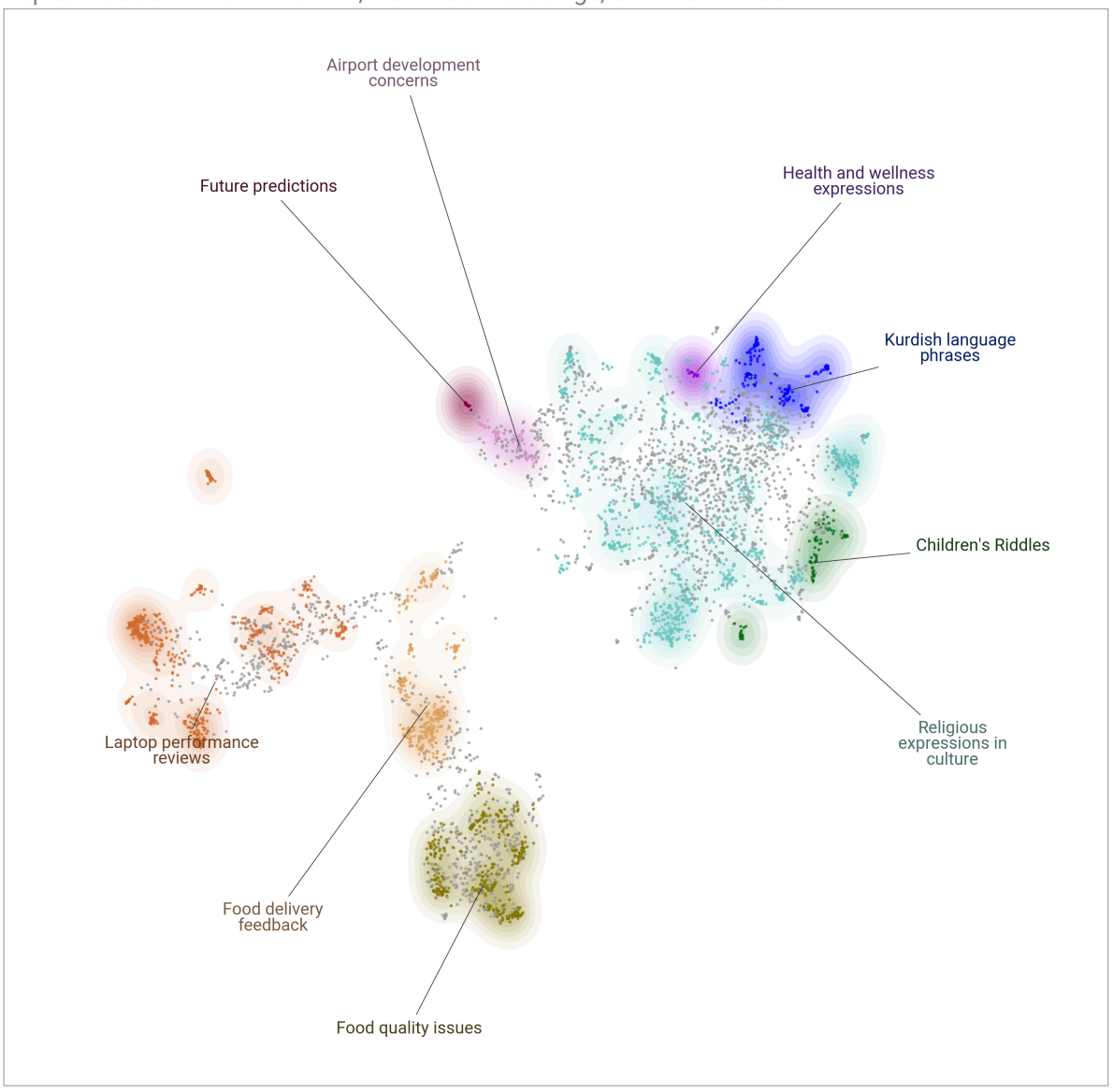

Figure 2: Topic modeling and visualization of Iranic languages and dialects. Semantic sentence embeddings are projected into a low-dimensional space using non-linear manifold learning and clustered via density-based methods. Automatically generated topic labels produced by a large language model enable interpretable analysis of cross-lingual and dialectal variation.