

Improving Persian Sentiment Classification through LLM-Assisted Label Refinement and Small Language Models

Maziyar Amini Alashti

Department of Computer Engineering
Imam Javad University College
Yazd, Iran
mzyramini@gmail.com

Seyed Abbas Bagheri

Department of Computer Engineering
Imam Javad University College
Yazd, Iran
sa.bagheri@iju.ir

Abstract

Persian sentiment analysis faces challenges in label quality and computationally efficient model adaptation. This paper investigates binary sentiment classification of Persian food-delivery reviews using small language models. We construct a 13,000-instance subset of the SnappFood dataset balanced according to LLM-refined labels and assess the refined labels through a stratified 500-example audit by two annotators ($\kappa = 0.972$). Four small language models (Gemma-3-1B, Gemma-3-270M, Llama-3.2-1B, and Qwen3-0.6B) are adapted via supervised instruction tuning with LoRA and 4-bit quantization on a single T4 GPU and compared with fine-tuned ParsBERT and XLM-RoBERTa-base baselines. Under the reported model-specific training configurations, all four fine-tuned generative models exceed 95% accuracy, with Llama-3.2-1B achieving the highest observed result (0.981 accuracy and 0.981 weighted F1) and a higher test-set accuracy than both encoder baselines. Label-refinement ablations, in which models are trained on original or refined labels and evaluated on the same refined test set, yield gains of 4.8–5.2 percentage points. These results support parameter-efficient instruction tuning and LLM-assisted label refinement as an effective approach for this Persian food-delivery benchmark under limited training resources, while generalization beyond this domain remains to be evaluated.

1 Introduction

Sentiment analysis is one of the central tasks in natural language processing (NLP), aiming to identify opinions, attitudes, and emotions expressed in text (Liu, 2012). It has many practical applications, including customer feedback analysis, social media monitoring, recommendation systems, market research, and public opinion analysis. While sentiment analysis has been studied extensively for English, Persian sentiment analysis remains more challenging due to the relatively limited availability

of high-quality annotated datasets, informal writing styles, orthographic variation, and domain-specific language use (Shamsfard, 2019).

In recent years, Persian user-generated content has grown rapidly across online stores, social media platforms, food-delivery services, and review websites. This growth has created new opportunities for training and evaluating sentiment analysis systems. However, many available Persian sentiment datasets contain noisy labels, especially when labels are derived automatically from ratings or user feedback signals. Label noise can reduce model reliability and may lead to misleading evaluation results. Therefore, dataset quality is an important factor in Persian sentiment classification.

Earlier work on Persian sentiment analysis explored lexicon-based approaches, traditional machine learning algorithms, and deep learning models such as CNN, LSTM, and BiLSTM architectures (Hajmohammadi and Ibrahim, 2013; Roshanfekr et al., 2017; Dashtipour et al., 2021). More recently, transformer-based models, especially ParsBERT, have achieved strong results in Persian language understanding and sentiment classification (Farahani et al., 2021; Zardak et al., 2023). However, most previous Persian sentiment analysis studies have focused on encoder-based models or task-specific neural architectures. In contrast, small language models have received less attention for Persian sentiment classification, particularly in an instruction-tuning setting.

Generative language models can be formulated as instruction-following classifiers by asking them to generate a sentiment label in response to a prompt. However, full fine-tuning of language models can be computationally expensive. Parameter-efficient fine-tuning methods such as LoRA allow models to be adapted by training only a small number of additional parameters (Hu et al., 2022). Combined with 4-bit quantization, LoRA makes it possible to fine-tune small language models on limited

hardware, such as a single T4 GPU.

In this paper, we investigate the effectiveness of small language models for binary Persian sentiment classification. We use a SnappFood Persian sentiment subset balanced according to refined labels and apply LLM-assisted label refinement to reduce noise in the original labels. We then evaluate four small generative language models and compare them with fine-tuned ParsBERT and XLM-RoBERTa-base encoder baselines.

The main contributions of this paper are as follows:

- We conduct a systematic evaluation of small generative language models adapted with instruction-formatted supervised fine-tuning for binary sentiment classification of Persian food-delivery reviews using a common dataset split and evaluation protocol.
- Under the reported model-specific training configurations, we show that parameter-efficient fine-tuning (LoRA + 4-bit quantization) enables the strongest small generative models to achieve higher test-set scores than the encoder baselines under the evaluated configurations while remaining trainable on a single T4 GPU.
- We construct a relabeled 13K subset of the SnappFood reviews dataset using LLM-assisted label refinement and assess its labels through a stratified 500-example, two-annotator manual audit ($\kappa = 0.972$), providing evidence that most audited label changes align with human judgments.
- We analyze LLM-assisted label refinement through controlled ablations, showing gains of 4.8–5.2 percentage points when models are trained with refined rather than original labels and evaluated against the refined test labels.

2 Related Work

Prior work relevant to this study falls into three main areas: Persian sentiment analysis, LLM-assisted data annotation, and instruction-based classification with generative language models.

2.1 Persian Sentiment Analysis

Persian sentiment analysis has been studied using lexicon-based methods, traditional machine learning, deep learning, and transformer-based models. Early work often relied on sentiment lexicons

and manually designed linguistic resources. [Amiri et al. \(2015\)](#) developed a lexicon-based Persian sentiment analysis system using a GATE pipeline, while [Akhoundzade and Devin \(2019\)](#) investigated unsupervised methods for expanding Persian sentiment lexicons. Although lexicon-based approaches are interpretable and require limited annotated data, they often struggle with context-dependent polarity, negation, sarcasm, and informal expressions.

Traditional supervised learning methods were later applied to Persian sentiment classification. [Hajmohammadi and Ibrahim \(2013\)](#) proposed an SVM-based method for Persian sentiment analysis, and [Saraee and Bagheri \(2013\)](#) studied feature selection methods using Naive Bayes for Persian opinion classification. [Basiri and Kabiri \(2017\)](#) introduced SPersent, a large sentence-level Persian sentiment dataset, contributing to the development of supervised sentiment models. These approaches improved over purely lexicon-based methods but typically depended on sparse features such as bag-of-words, n-grams, and manually engineered representations.

Deep learning methods reduced the need for manual feature engineering by learning distributed representations automatically. [Roshanfekr et al. \(2017\)](#) applied deep neural models to Persian sentiment analysis, while [Dashtipour et al. \(2021\)](#) investigated Persian movie review sentiment classification using fastText embeddings and neural architectures including CNN, LSTM, and BiLSTM models. [Dastgheib et al. \(2020\)](#) explored deep learning for Persian document sentiment analysis, and [Shumaly et al. \(2021\)](#) showed that CNN models with fastText embeddings can achieve strong performance on Persian online store reviews.

More recently, pre-trained transformer-based models have become strong baselines for Persian NLP. [Farahani et al. \(2021\)](#) introduced ParsBERT, a Persian transformer-based language model that has been widely used for Persian language understanding tasks. [Karimi and Shahrabadi \(2019\)](#) explored BERT-based sentiment analysis for Persian texts, and [Zardak et al. \(2023\)](#) used BERT-based neural models for Persian sentiment classification. In addition to Persian-specific encoders, multilingual models such as XLM-RoBERTa ([Conneau et al., 2020](#)) provide strong cross-lingual representations and are commonly used as baselines for non-English NLP tasks. These works establish encoder-based transformers as strong baselines for Persian sentiment analysis, but they do not directly examine

small generative language models in an instruction-tuning setting.

2.2 LLM-Assisted Data Annotation

Recent studies have investigated the use of large language models as annotators or label refiners. Wang et al. (2021) explored GPT-3 as a low-cost data labeler for NLP tasks and showed that LLM-generated labels can reduce annotation costs while supporting downstream model training. Gilardi et al. (2023) compared ChatGPT with crowd workers for text annotation tasks and found that ChatGPT can achieve competitive or superior performance in several annotation settings. He et al. (2024) proposed AnnoLLM, an LLM-powered annotation framework that uses an explain-then-annotate strategy to improve annotation quality.

These studies suggest that LLMs can be useful for scalable annotation, especially when human labeling is expensive or when existing labels are noisy. However, LLM-generated labels are not equivalent to expert gold-standard annotations and require validation. In this work, we use LLM-assisted label refinement not as a replacement for human evaluation, but as a noise-reduction step applied to a subset of an existing Persian sentiment dataset. We further validate the refined labels through a stratified two-annotator manual audit.

2.3 Instruction Tuning for Classification

Generative language models can be adapted to classification tasks by formulating the task as instruction following: the model receives an input text and generates a label as output. Instruction tuning has been shown to improve the ability of language models to follow task descriptions and generalize across NLP tasks (Wei et al., 2022; Chung et al., 2024). For text classification, Sun et al. (2023) investigated LLM-based classification and showed that prompting strategies can improve performance on classification benchmarks, while also noting that LLMs may struggle with complex linguistic phenomena such as contrast and irony. In sentiment analysis, Zhang et al. (2023) showed that converting supervised financial sentiment data into instruction-style data and fine-tuning a general-purpose LLM can improve financial sentiment classification.

This line of work motivates the use of generative models as instruction-following classifiers. However, most existing studies focus on English or high-resource domains, and many use relatively large models. To the best of our knowledge, the use of

instruction-tuned small generative language models for Persian sentiment classification has received little direct investigation in the existing literature.

2.4 Position of This Work

Prior Persian sentiment analysis research has primarily focused on lexicon-based methods, traditional machine learning, deep neural networks, and encoder-based transformers. Separately, recent studies have explored LLM-assisted annotation and instruction tuning for classification tasks. This work brings these directions together by evaluating instruction-tuned small language models for Persian sentiment classification on an LLM-refined dataset and comparing them with strong encoder-based baselines using parameter-efficient fine-tuning.

3 Dataset and Label Refinement

The experiments were conducted using a relabeled subset of the publicly available SnappFood Persian sentiment dataset (Tehranipour, 2022). The dataset contains Persian food-delivery user reviews labeled with binary sentiment classes. The public dataset is balanced across positive and negative review labels.

During preliminary inspection, we observed examples whose original labels contradicted the sentiment expressed in the review text. Since label noise can negatively affect sentiment classification, we applied an LLM-assisted label refinement procedure. We first randomly sampled 6,500 reviews originally labeled positive and 6,500 reviews originally labeled negative from the SnappFood dataset. After refinement, the resulting set was no longer balanced, with more examples assigned to the negative class than to the positive class. We therefore sampled additional reviews from the remaining original dataset and applied the same refinement procedure to these examples until sufficient refined-positive and refined-negative examples were available. The final benchmark was then constructed by retaining 6,500 refined-positive and 6,500 refined-negative reviews. Thus, the final 13,000-example subset is balanced with respect to the refined labels, not necessarily with respect to the original SnappFood labels.

Relabeling was performed through the OpenAI API using the gpt-5-mini alias, accessed on 2025-08-26. Temperature was not explicitly specified in the retained API code. No immutable snapshot identifier was available in the retained API records.

The same prompt template was used for all samples. The model was instructed to determine the dominant sentiment and return a JSON object containing one of two labels: positive or negative. The exact relabeling prompt and code are included in the released artifacts. The returned JSON was parsed programmatically, and all retained examples were successfully parsed into one of the two permitted sentiment labels.

Among the final retained 13,000 examples, the refined label differed from the original SnappFood label in 1,635 cases, corresponding to 12.6% of the retained dataset. To assess label quality, we conducted a stratified manual audit of 500 examples: 250 of the 1,635 changed-label examples and 250 of the 11,365 unchanged-label examples. Two native Persian speakers independently assigned each audited review a binary label according to its dominant expressed sentiment. Annotators were not shown the original or refined labels during annotation.

Table 1 reports each annotator’s agreement with both the refined and original labels. In the changed-label stratum, the refined labels agreed with Annotators 1 and 2 in 242 of 250 cases (96.8%) and 244 of 250 cases (97.6%), respectively, whereas the original labels agreed in 8 cases (3.2%) and 6 cases (2.4%). In the unchanged-label stratum, the original and refined labels are identical and therefore have the same agreement values. Thus, most label changes in the audited stratum aligned with the annotators’ judgments rather than with the original labels. This conclusion is limited to the sampled cases and does not establish the correctness of every one of the 1,635 changed labels.

The two annotators agreed on 493 of the 500 audited examples, corresponding to 98.6% raw agreement and Cohen’s $\kappa = 0.972$. This indicates high consistency between the annotators on the audited sample. Because the audit covered only a stratified subset and was not followed by full expert adjudication, we treat the refined dataset as a silver-standard resource.

The final dataset was balanced according to the refined labels and split into 10,000 training examples, 1,000 validation examples, and 2,000 test examples. No manual text normalization, stopword removal, stemming, lemmatization, or feature engineering was applied. This decision was made because transformer-based models use subword tokenization and can often benefit from preserving the original surface form of user-

generated text (Ho, 2025). To support reproducibility, the relabeled 13,000-instance dataset and fixed train/validation/test splits are available on the Hugging Face Hub (Amini Alashti and Bagheri, 2026b). The label-refinement prompt and refinement code are available in the accompanying GitHub repository (Amini Alashti and Bagheri, 2026a).

4 Methodology

We formulate binary Persian sentiment classification as an instruction-following task for generative language models. Each review is inserted into a fixed prompt template, and the model is trained to generate one of two sentiment labels: positive or negative. The prompt template used for the generative models is shown below:

```
What is the sentiment of the Persian
text? Respond only with positive or
negative based on the dominant sentiment.
Even if one sentiment is slightly
present, choose the label based on the
stronger tone.
Content: [REVIEW]
Sentiment:
```

We evaluate four small generative language models: google/gemma-3-1b-it and google/gemma-3-270m-it from the Gemma 3 family (Gemma Team et al., 2025), meta-llama/Llama-3.2-1B from the Llama 3.2 family (Meta AI, 2024), and Qwen/Qwen3-0.6B from the Qwen3 family (Qwen Team, 2025). These models were fine-tuned using supervised instruction tuning with LoRA (Hu et al., 2022) and 4-bit quantization, following the general motivation of quantized parameter-efficient fine-tuning (Dettmers et al., 2023). This setup was selected to reduce GPU memory usage and enable training in a limited-resource environment.

For comparison, we also fine-tuned two encoder-based baselines: HooshvareLab/bert-base-parsbert-uncased, a Persian BERT-based encoder model (Farahani et al., 2021), and xlm-roberta-base, a multilingual transformer encoder (Conneau et al., 2020). Since these models are not generative instruction-following models, each encoder was fine-tuned with a classification head rather than evaluated through label generation.

All models were trained using Hugging Face Transformers and PyTorch, with TRL SFTTrainer, PEFT, and bitsandbytes additionally used for the generative models. The generative runs used a

Audit subset	Compared label	A1 agreement	A2 agreement	Mean agreement
Changed ($N = 250$)	Refined	242 (96.8%)	244 (97.6%)	97.2%
	Original	8 (3.2%)	6 (2.4%)	2.8%
Unchanged ($N = 250$)	Refined/original	247 (98.8%)	246 (98.4%)	98.6%
Stratified total ($N = 500$)	Refined	489 (97.8%)	490 (98.0%)	97.9%
	Original	255 (51.0%)	252 (50.4%)	50.7%

Table 1: Human agreement with the refined and original labels in the stratified audit. Mean agreement is the arithmetic mean of the A1 and A2 agreement percentages. Original and refined labels are identical in the unchanged stratum. Because changed and unchanged examples were sampled equally, the stratified-total rows describe the audited sample and should not be interpreted as corpus-wide estimates of label accuracy.

Split	Positive	Negative	Total
Train	5,000	5,000	10,000
Validation	500	500	1,000
Test	1,000	1,000	2,000
Total	6,500	6,500	13,000

Table 2: Dataset statistics for the final retained subset after label refinement. Counts are based on the refined labels.

per-device batch size of 1 with eight gradient-accumulation steps, while the encoder runs used a batch size of 8 without gradient accumulation, yielding an effective training batch size of 8 for both model classes. Other optimization settings, including learning rates, schedulers, epoch counts, and fine-tuning strategies, were model-class-specific rather than compute-matched. The comparison therefore evaluates the reported practical fine-tuning recipes on a common split and evaluation set rather than isolating the effect of model architecture. The main configurations are shown in Table 3.

During inference, deterministic decoding was used for all generative models with `do_sample=False`, `temperature=0.0`, and `max_new_tokens=5`. The generated output was then mapped to one of the two sentiment labels. To assess robustness to random initialization, we repeated the main fine-tuning runs for ParsBERT, Gemma-3-1B, and Llama-3.2-1B using two additional random seeds (123 and 456). Seed 42 was the default seed used in the original experiments and corresponds to the results reported in the main tables.

5 Experiments and Results

We evaluated all models on the same balanced test set of 2,000 examples, with 1,000 examples per class. Accuracy, precision, recall, and F1-score

were used as evaluation metrics. Precision, recall, and F1-score are reported as support-weighted averages. Because the two classes have equal support in the test set, weighted F1 is mathematically identical to macro F1. The reported F1 values therefore represent both metrics.

5.1 Pre-Fine-Tuning Performance

Before fine-tuning, the small language models showed varying performance when prompted directly for sentiment classification. Gemma-3-1B and Llama-3.2-1B achieved pre-fine-tuning accuracies of 85.8% and 81.1%, respectively, while Gemma-3-270M and Qwen3-0.6B achieved 52.1% and 55.3%, respectively, close to chance performance on the balanced test set. These results indicate that prompting alone did not provide reliable classification performance across all evaluated models on this Persian food-delivery dataset.

ParsBERT and XLM-RoBERTa-base were not included in the pre-fine-tuning prompting comparison because they are encoder-only models and they do not directly generate sentiment labels without an added classification head. Table 4 shows the accuracy before and after fine-tuning for the generative models.

5.2 Fine-Tuned Model Performance

After fine-tuning, all six models achieved accuracy above 95%. Llama-3.2-1B achieved the highest observed performance, with 0.981 accuracy and 0.981 F1, followed by Gemma-3-1B with 0.977 accuracy. Among the encoder baselines, XLM-RoBERTa-base achieved 0.967 accuracy and 0.967 F1, slightly above ParsBERT at 0.964 accuracy and 0.963 F1. Under the reported model-specific training configurations, Llama-3.2-1B and Gemma-3-1B therefore achieved higher point-estimate accuracy than both encoder baselines. These results indicate the competitiveness of the two 1B genera-

Parameter	Generative Models	Encoder Baselines
Models	Gemma, Llama, Qwen	ParsBERT, XLM-RoBERTa-base
Fine-tuning method	LoRA instruction tuning	Full classifier fine-tuning
Quantization	4-bit	None
LoRA rank	64	–
LoRA alpha	32	–
LoRA dropout	0.05	–
Target modules	all-linear	–
Train batch size	1	8
Eval batch size	1	8
Gradient accumulation steps	8	1
Effective train batch size	8	8
Learning rate	2e-4	2e-5
Epochs	2	10
LR scheduler	cosine	linear
Warmup steps	500	–
Optimizer	paged_adamw_32bit	adamw_torch
Adam betas	–	(0.9, 0.98)
Adam epsilon	–	1e-8
Precision	fp16	mixed precision
Maximum sequence length	512	512
Evaluation strategy	every 112 steps	every epoch
Save strategy	–	every epoch
Best model criterion	–	eval_accuracy
Hardware	Google Colab T4 GPU	Google Colab T4 GPU

Table 3: Fine-tuning configuration for generative models and encoder baselines.

Model	Before (%)	After (%)	Gain (pp)
Gemma-3-1B	85.8	97.7	+11.9
Gemma-3-270M	52.1	95.7	+43.6
Llama-3.2-1B	81.1	98.1	+17.0
Qwen3-0.6B	55.3	96.6	+41.3

Table 4: Accuracy before and after fine-tuning for the generative models. Before and after values are percentages, gains are percentage points.

tive models under the evaluated practical recipes, but do not by themselves establish an inherent advantage of generative over encoder architectures. The full results are shown in Table 5.

5.3 Robustness Across Random Seeds

To assess sensitivity to random initialization, we repeated the fine-tuning runs for ParsBERT, Gemma-3-1B, and Llama-3.2-1B using two additional random seeds (123 and 456) while keeping the dataset split fixed. Seed 42 was used for the results in the main tables. Table 6 summarizes accuracy across the three seeds.

Performance was highly stable for these three models. Gemma-3-1B produced the same accuracy in all three runs, while ParsBERT and Llama-3.2-1B varied by at most 0.1 accuracy points. Their ranking remained unchanged across the three seeds, indicating that the reported comparison among these models is not sensitive to random initial-

Model	Acc.	Prec.	Rec.	F1
<i>Encoder baselines</i>				
ParsBERT	0.964	0.964	0.964	0.963
XLM-RoBERTa-base	0.967	0.967	0.967	0.967
<i>Generative models</i>				
Gemma-3-270M	0.957	0.957	0.957	0.956
Qwen3-0.6B	0.966	0.966	0.966	0.966
Gemma-3-1B	0.977	0.977	0.977	0.976
Llama-3.2-1B	0.981	0.981	0.981	0.981

Table 5: Fine-tuned model performance on the test set. Precision, recall, and F1 are support-weighted averages. Because the test set contains 1,000 examples per class, weighted and macro F1 are identical.

Model	Accuracy (Mean $\pm$ SD)
Gemma-3-1B	0.9770 $\pm$ 0.0000
Llama-3.2-1B	0.9813 $\pm$ 0.0006
ParsBERT	0.9637 $\pm$ 0.0006

Table 6: Accuracy across three random seeds on the refined test set.

ization. This analysis does not measure variation across dataset samples or train/validation/test splits.

5.4 Effect of Label Refinement

To assess the downstream effect of label refinement, we conducted an ablation using one encoder-based model and two generative models. Gemma-3-1B, Llama-3.2-1B, and ParsBERT were fine-tuned on

Model	Train labels	Acc.	Gain
Gemma-3-1B	Original	0.929	–
Gemma-3-1B	Refined	0.977	+0.048
Llama-3.2-1B	Original	0.931	–
Llama-3.2-1B	Refined	0.981	+0.050
ParsBERT	Original	0.912	–
ParsBERT	Refined	0.964	+0.052

Table 7: Effect of label refinement. Models trained with original and refined labels use the same data split and are evaluated on the same refined test set.

Model	TN	FP	FN	TP	Err.
ParsBERT	966	34	39	961	73
XLNet-RoBERTa-base	965	35	31	969	66
Gemma-3-270M	950	50	37	963	87
Qwen3-0.6B	971	29	40	960	69
Gemma-3-1B	982	18	29	971	47
Llama-3.2-1B	977	23	15	985	38

Table 8: Confusion matrix results on the test set.

the same training examples using either the original or refined training labels, and all variants were evaluated against the same refined test labels. As shown in Table 7, Gemma-3-1B improved from 0.929 to 0.977 accuracy (+4.8 percentage points), Llama-3.2-1B improved from 0.931 to 0.981 (+5.0 points), and ParsBERT improved from 0.912 to 0.964 (+5.2 points). These consistent gains show that the refined training labels are more effective supervision for predicting the refined evaluation target. Together with the manual audit, the results are consistent with the interpretation that refinement removed substantial label noise, although the ablation alone does not independently establish the correctness of every refined test label.

5.5 Confusion Matrix Analysis

Table 8 reports the confusion matrix values for each fine-tuned model. Llama-3.2-1B produced the fewest total errors, with 38 misclassified examples out of 2,000. Gemma-3-1B produced 47 errors. Among encoder baselines, XLNet-RoBERTa-base produced 66 errors, while ParsBERT produced 73 errors. The results indicate that Llama-3.2-1B not only achieved the highest overall score but also showed balanced performance across positive and negative classes.

5.6 Error Analysis

To better understand the remaining model failures, we manually inspected the errors made by the best-

Error type	Count
Mixed sentiment / aspect conflict	26
Sarcasm or irony	2
Short or ambiguous text	2
Implicit or polite complaint	5
Other	3

Table 9: Qualitative error categories for the 38 misclassified examples produced by Llama-3.2-1B.

performing model, Llama-3.2-1B. The analysis showed that most errors were not simple polarity mistakes, but involved linguistically or pragmatically difficult cases. The most frequent error type was mixed sentiment or aspect conflict, where a review expressed satisfaction with one aspect of the order while criticizing another. For example, a review may praise the taste of the food while complaining that it arrived cold or late. Such cases are difficult for binary review-level classification because the dominant sentiment is sometimes ambiguous.

Other errors involved sarcastic or ironic complaints, short or ambiguous reviews, and implicit or polite complaints. Sarcastic examples often contained positive words but conveyed a negative overall meaning through context. Short reviews such as “It was ordinary” provided limited evidence for assigning a clear positive or negative label. Polite complaints were also challenging because expressions of gratitude or politeness sometimes appeared together with negative feedback. These observations suggest that the remaining errors are partly caused by the limitations of binary review-level sentiment classification. Persian food-delivery reviews often contain multiple aspects, such as food quality, delivery time, packaging, and restaurant behavior, which may not be fully captured by a single sentiment label.

5.7 Discussion

The results show that parameter-efficient instruction tuning substantially improves small language models for Persian sentiment classification. Before fine-tuning, model performance varied considerably. Gemma-3-1B and Llama-3.2-1B achieved reasonable accuracy, but Gemma-3-270M and Qwen3-0.6B performed near random classification. After fine-tuning, however, all generative models achieved strong performance, with accuracy above 95%.

The largest improvements were observed for

Gemma-3-270M and Qwen3-0.6B, which improved by 43.6 and 41.3 percentage points, respectively. This suggests that small language models may have limited zero-shot reliability for Persian sentiment classification, but they can become highly effective after supervised instruction tuning.

On the evaluated SnappFood split, the two 1B generative models were competitive with the Persian-specific and multilingual encoder baselines. XLM-RoBERTa-base achieved 0.967 accuracy, slightly above ParsBERT at 0.964, while Llama-3.2-1B and Gemma-3-1B achieved higher test-set accuracies (0.981 and 0.977, respectively). These results show that parameter-efficient instruction tuning is a viable alternative to encoder-based classification for this in-domain task under the reported training configurations. Because the optimization recipes were model-class-specific, the results should not be interpreted as establishing an inherent architectural advantage for generative models.

The use of 4-bit quantization and LoRA enabled all four generative models to be fine-tuned on a single Google Colab T4 GPU. This demonstrates the feasibility of the adaptation procedure under limited training hardware, although the broader pipeline also depends on a hosted LLM for label refinement.

The label-refinement ablation further supports the importance of annotation quality on this dataset. Training with refined rather than original labels increased accuracy by 4.8 points for Gemma-3-1B, 5.0 points for Llama-3.2-1B, and 5.2 points for ParsBERT (Table 7). These consistent gains across encoder-based and generative models indicate that the refined labels provide a more effective supervision signal for the refined evaluation target.

From a teacher–student perspective, the pipeline can be viewed as a form of offline hard-label knowledge transfer related to knowledge distillation (Hinton et al., 2015). GPT-5-mini assigns a binary label to each review during dataset construction, and the smaller models subsequently learn from these labels without access to the teacher at inference time. Unlike conventional knowledge distillation, the procedure does not transfer soft probability distributions, intermediate representations, or teacher rationales. It is therefore best characterized as a hybrid of LLM-assisted label refinement and hard-label teacher–student transfer rather than conventional distillation. The label-blind manual audit supports the noise-correction interpretation for the

audited examples, while the ablation reflects the combined effect of noise reduction and alignment with the refined labeling rule.

The error analysis further shows that the remaining mistakes are concentrated in cases where binary review-level sentiment is inherently limited. Most errors involved mixed sentiment or aspect conflict, such as reviews that praised food quality while criticizing delivery time or temperature. This suggests that further improvements may require moving beyond binary classification toward aspect-based or multi-label sentiment analysis, especially for service-review domains where users frequently evaluate multiple aspects in a single comment.

6 Conclusion

This paper examined whether small language models can be adapted effectively for binary sentiment classification of Persian food-delivery reviews under limited training resources. Using a SnappFood subset balanced according to refined labels, we compared four generative models with ParsBERT and XLM-RoBERTa-base encoder baselines. Parameter-efficient instruction tuning with LoRA and 4-bit quantization produced accuracy above 95% for all four generative models. Under the reported model-specific configurations, Llama-3.2-1B achieved the highest observed accuracy and a higher point estimate than both encoder baselines, while the smaller generative models also improved substantially after fine-tuning.

The LLM-assisted refinement process changed 12.6% of the selected labels. A stratified 500-example audit by two annotators showed that, within the audited changed-label stratum, refined labels agreed much more closely with human judgments than the original labels, with high inter-annotator agreement ($\kappa = 0.972$). Training with refined rather than original labels improved accuracy by 4.8 percentage points for Gemma-3-1B, 5.0 points for Llama-3.2-1B, and 5.2 points for ParsBERT (Table 7).

Additional runs for ParsBERT, Gemma-3-1B, and Llama-3.2-1B were stable across three random seeds on the fixed data split. This supports initialization stability for these three models but does not assess variation across alternative dataset samples or train/test splits.

Future work should extend this study in several directions. First, broader baseline comparisons with multilingual encoders and classical classifiers

should be conducted across multiple Persian sentiment datasets. Second, larger human validation with multiple annotators is needed to better assess the quality of LLM-refined labels. Third, future models should be evaluated in cross-domain settings, such as social media, product reviews, and news comments. Finally, aspect-based and multi-label sentiment formulations may better capture Persian reviews that contain both positive and negative opinions.

Limitations

First, this study evaluates one 13,000-review subset of SnappFood using one fixed train/validation/test split. Food-delivery reviews may contain more explicit polarity cues than domains with highly informal, implicit, or context-dependent sentiment, such as social media and news comments. Accordingly, the reported results establish in-domain performance on this benchmark and should not be assumed to generalize to Persian sentiment classification broadly. Cross-domain and multiple-split evaluation remains future work.

Second, the refined labels remain silver-standard rather than gold-standard annotations. The stratified audit covered 500 examples, including 250 of the 1,635 changed-label cases, and deliberately oversampled changed examples. Its pooled agreement values therefore characterize the audited sample rather than providing corpus-wide estimates of label accuracy. Although the two annotators showed high agreement, the audit was not followed by full expert adjudication and cannot establish the correctness of every refined label. In addition, the label-refinement ablation was evaluated against test labels produced by the same refinement procedure.

The use of a proprietary hosted model for label refinement also limits exact reproducibility. Even when the prompt, API model identifier, and access date are reported, provider-side model updates or retirement may prevent future users from reproducing the same outputs exactly. Releasing the refined labels, label-refinement prompt, and refinement code will facilitate reproduction of the annotation procedure and enable further research using the refined dataset, even if the original API-based refinement cannot be regenerated exactly.

Third, the experiments focus only on binary sentiment classification and do not include neutral, mixed, or aspect-level sentiment categories. Many real-world reviews contain both positive

and negative opinions, and a binary label may not fully represent their sentiment structure. Fourth, the additional-seed experiments cover ParsBERT, Gemma-3-1B, and Llama-3.2-1B only and assess sensitivity to random initialization on a fixed data split. They do not measure robustness across alternative samples, splits, hyperparameters, or domains.

Finally, the generative and encoder models were trained with model-class-specific optimization schedules and epoch counts rather than through a compute-matched hyperparameter sweep. Although their effective training batch sizes were the same, the comparison should be interpreted as an evaluation of the reported practical recipes rather than as an architecture-controlled comparison. Future work could complement point-estimate comparisons with paired significance tests.

References

- Reza Akhoundzade and Kamran Hajiabadi Devin. 2019. Persian sentiment lexicon expansion using unsupervised learning methods. In *2019 9th International Conference on Computer and Knowledge Engineering*, pages 461–465. IEEE.
- Maziyar Amini Alashti and Seyed Abbas Bagheri. 2026a. [Code and prompts for LLM-assisted label refinement of Persian sentiment data](#). GitHub repository.
- Maziyar Amini Alashti and Seyed Abbas Bagheri. 2026b. [Refined SnappFood Persian sentiment dataset](#). Hugging Face dataset.
- Fatemeh Amiri, Simon Scerri, and Mehdi Khodashahi. 2015. Lexicon-based sentiment analysis for persian text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 9–16.
- Mohammad Ehsan Basiri and Arman Kabiri. 2017. Sentence-level sentiment analysis in persian. In *2017 3rd International Conference on Pattern Recognition and Image Analysis*, pages 84–89. IEEE.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- 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.
- Kia Dashtipour, Mandar Gogate, Ahsan Adeel, Hadi Larijani, and Amir Hussain. 2021. Sentiment analysis of persian movie reviews using deep learning. *Entropy*, 23(5):596.
- Mohammad Bagher Dastgheib, Saeedeh Koleini, and Fariba Rasti. 2020. The application of deep learning in persian documents sentiment analysis. *International Journal of Information Science and Management*, 18(1):1–15.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*.
- Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and M Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.
- 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](#). Preprint, arXiv:2503.19786.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Mohammad Sadegh Hajmohammadi and Roliana Ibrahim. 2013. A svm-based method for sentiment analysis in persian language. In *International Conference on Graphic and Image Processing*, volume 8768, pages 697–701. SPIE.
- Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Chuk Fong Ho. 2025. [The declining role of lemmatization and stemming in modern nlp pipelines](#). In *2025 9th International Conference on Man-Machine Systems (ICoMMS)*, pages 426–431.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Sahar Karimi and Fatemeh Sadat Shahrabadi. 2019. Sentiment analysis using bert and deep learning on persian texts. *Technology of Deep Learning*.
- Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Meta AI. 2024. [Llama 3.2 model card](#). Model card. Accessed: 2026-06-05.
- Qwen Team. 2025. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.
- Behnam Roshanfekr, Shahram Khadivi, and Mohammad Rahmati. 2017. Sentiment analysis using deep learning on persian texts. In *2017 Iranian Conference on Electrical Engineering*, pages 1503–1508. IEEE.
- Mohamad Saraee and Ayoub Bagheri. 2013. Feature selection methods in persian sentiment analysis. In *International Conference on Application of Natural Language to Information Systems*, pages 303–308. Springer.
- Mehrnoush Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *International Conference Language Technologies for All*.
- Sara Shumaly, Mohammad Yazdinejad, and Yike Guo. 2021. Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fasttext embeddings. *PeerJ Computer Science*, 7:e422.
- Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Tianwei Guo, and Yang Zhang. 2023. [Text classification via large language models](#). Preprint, arXiv:2305.08377.
- Soheil Tehranipour. 2022. [Snappfood persian sentiment analysis dataset](#). Kaggle dataset.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.
- Seyed Reza Zardak, Ali Hossein Rasekh, and Mohammad Sadegh Bashkari. 2023. Persian text sentiment analysis based on bert and neural networks. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 47(4):1623–1634.
- Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. [Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models](#). Preprint, arXiv:2306.12659.