Farsi Review Rating Prediction Using Prompt Engineering and RAG

Vanooshe Nazari[*]**, Alireza Moradi**[*]**, Sauleh Eetemadi, Mohammad Fathian**
Iran University of Science and Technology
V_nazari@ind.iust.ac.ir,a_moradi78@ind.iust.ac.ir,sauleh@iust.ac.ir
fathian@iust.ac.ir

Abstract

Review rating prediction is a crucial task that benefits both businesses and customers by enhancing decision-making and improving service quality. In this study, we propose several prompt-based methods for predicting rating of reviews using large language models, specifically GPT-3.5 Turbo and Dorna. We utilize BaSalam dataset, sourced from an Iranian online marketplace. Our approach includes zero-shot and few-shot prompting, as well as Retrieval-Augmented Generation. We evaluate the effectiveness of our methods by comparing them to baseline models, demonstrating superior performance in terms of mean absolute error (MAE) and mean squared error (MSE).

1 Introduction and Related Works

Review rating prediction benefits consumers and businesses by enhancing user experience through better recommendations. Businesses can improve customer service by understanding review ratings.

This study explores methods for predicting ratings for review texts, a task similar to sentiment analysis and framed as a regression problem predicting ratings from 1 to 5. We focus on prompting techniques like retrieval-augmented generation (RAG), few-shot, and zero-shot prompting. We also use Support Vector Machines(SVM) (Tsochantaridis et al., 2004) and K-Nearest Neighbors(KNN) (Cover and Hart, 1967) as baseline models. Our study leverages GPT-3.5 Turbo (Brown et al., 2020) and Dorna (PartAI, 2024). Dorna is an open-source Farsi large language model, chosen for its strong performance on various benchmarks. We also use Multilingual E5 base (Wang et al., 2023) for embeddings. The dataset comprises review texts sourced from BaSalam, an Iranian online platform.

Several studies have explored methods for predicting review ratings. Bentaleb and Abouchabaka (2023) used transfer learning with XLNet and a two-phase fine-tuning method on the Yelp Dataset. Guda et al. (2022) proposed a Multi-tasked Joint BERT model for sentiment prediction in restaurant reviews. Liu (2020) used various machine learning and transformer-based models for restaurant ratings. Ahmed and Ghabayen (2022) proposed a deep learning framework using a Bi-GRU model for polarity and rating prediction. Other studies have used different feature sets and machine learning methods like SVM, Naive Bayes, and Logistic Regression (Elkouri, 2015; Asghar, 2016; Fan and Khademi, 2014). However, there has been no research on Farsi language employing RAG for review rating prediction, which motivates this study. Additionally, we utilize a few-shot prompting technique that incorporates users' historical reviews in the prompt, a method that has not yet been applied in Farsi review rating prediction.

2 Methodology

In this section, we introduce the dataset and outline the preprocessing steps, followed by a detailed description of the various proposed methods.

2.1 Dataset

The dataset utilized in this research is from the Iranian marketplace BaSalam (RadeAI, 2024). While the dataset includes several features, our analysis focuses on reviews and their corresponding ratings. The preprocessing steps include removing empty reviews and duplicates, and extracting a balanced subset for analysis using undersampling.

2.2 Proposed Methods

In this section, we provide a detailed overview of the various methods employed in this study. We utilize GPT-3.5 Turbo and Dorna as our language models, with all prompts formulated in Farsi. As baseline methods, we use SVM and KNN, using multilingual E5 base embeddings as their features.

[*] Equal contribution.

2.2.1 Zero-Shot Prompting

We employ zero-shot prompting, where the model is not provided with any examples. Instead, it is tasked solely with predicting the rating for a specific review.

2.2.2 Few-Shot Prompting

In this method, we utilize two distinct types of few-shot prompting. The first approach involves providing the model with fixed examples of each class along with their corresponding ratings. The second approach focuses on the user's past reviews; specifically, we supply the model with examples of each available rating class, chosen randomly from the same user's previous reviews. This allows the model to learn the user's rating behavior and preferences more effectively.

2.2.3 Retrieval-Augmented Generation(RAG)

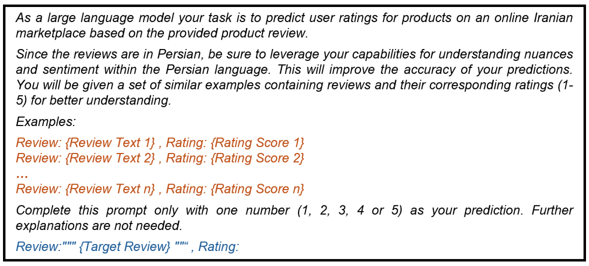
In this method, we retrieve the most similar examples to the target review for which we aim to predict the rating. For the retrieval process, we utilize the multilingual E5 base to generate embeddings, allowing us to search for the most similar examples. These examples are incorporated into the prompt to provide relevant context for predicting the rating through in-context learning.

Models	Method	MAE	MSE
Baselines	SVM	0.790	1.110
	KNN	0.760	1.048
GPT-3.5	Zero-Shot	0.632	0.924
	Fixed Example	0.603	0.851
	User-Specific Examples	0.602	**0.834**
	RAG with Two Examples	0.601	0.893
	RAG with Four Examples	0.589	0.867
	RAG with Eight Examples	**0.580**	0.864
Dorna	Zero-Shot	0.729	0.943
	Fixed Example	0.721	1.061
	User-Specific Examples	0.679	1.037
	RAG with Two Examples	0.727	1.145
	RAG with Four Examples	0.649	0.995
	RAG with Eight Examples	0.614	0.870

Table 1: Comparison of MAE and MSE for various prediction methods.

Language	Method	MAE	MSE
Farsi	Zero-Shot	0.632	0.924
	User-Specific Examples	0.602	**0.834**
	RAG with Eight Examples	**0.580**	0.864
Mixed	Zero-Shot	0.651	1.053
	User-Specific Examples	0.643	0.929
	RAG with Eight Examples	0.611	0.949

Table 2: Comparison of GPT-3.5 with different language settings (Only best methods of few-shot prompting and RAG are included).



As a large language model your task is to predict user ratings for products on an online Iranian marketplace based on the provided product review.

Since the reviews are in Persian, be sure to leverage your capabilities for understanding nuances and sentiment within the Persian language. This will improve the accuracy of your predictions. You will be given a set of similar examples containing reviews and their corresponding ratings (1-5) for better understanding.

Examples:

Review: {Review Text 1} , Rating: {Rating Score 1}
Review: {Review Text 2} , Rating: {Rating Score 2}
...
Review: {Review Text n} , Rating: {Rating Score n}

Complete this prompt only with one number (1, 2, 3, 4 or 5) as your prediction. Further explanations are not needed.

Review:""" {Target Review} ""“ , Rating:

Figure 1: A prompt used in the mixed-language setting.

3 Results

We conduct multiple experiments to predict review ratings using our dataset. Due to the ordinal nature of these ratings, we used regression metrics, specifically MAE and MSE to evaluate our methods.

Initially, we employ RAG with two, four, and eight examples. As shown in Table 1, performance consistently improves with the use of more examples, ultimately achieving the lowest MAE and MSE when employing eight examples, surpassing other RAG methods. Notably, RAG with eight examples also have the lowest MAE among all methods. However, we did not use more than eight examples due to cost constraints.

Moreover, we explore two methods for few-shot prompting. Table 1 shows that using user-specific examples yields better results. This improvement is attributed to the unique rating behaviors of individuals. The effectiveness of this method is evident in Table 1, which shows the lowest MSE.

We also conducted another experiment to test GPT-3.5 in different language settings. The results, presented in Table 2, compare two settings:

1. Farsi: The entire prompt is in Farsi.

2. Mixed-language: The task description is in English, but the examples are in Farsi. An example of these prompts can be seen in fig 1.

The Farsi setting outperforms the mixed-language setting, indicating that when the examples are in Farsi, it is more effective for the task description to be in Farsi as well.

4 Conclusion

In this study, we explored various prompting techniques for predicting review ratings in Farsi. The effectiveness of these methods is clear, as they significantly surpass baseline models. Notably, the GPT-3.5 model with eight retrieved examples, achieves the best performance in terms of MAE. Additionally, when we use user-specific examples, this methods outperforms others in terms of MSE.

References

Basem H Ahmed and Ayman S Ghabayen. 2022. Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, 13(7):3423–3432.

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Asmae Bentaleb and Jaafar Abouchabaka. 2023. Fine-tuning deep learning model for review rating prediction. *International Journal of Computing and Digital Systems*, 14(1):1061–1053.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Andrew Elkouri. 2015. Predicting the sentiment polarity and rating of yelp reviews. *arXiv preprint arXiv:1512.06303*.

Mingming Fan and Maryam Khademi. 2014. Predicting a business star in yelp from its reviews text alone. *arXiv preprint arXiv:1401.0864*.

Bhanu Prakash Reddy Guda, Mashrin Srivastava, and Deep Karkhanis. 2022. Sentiment analysis: Predicting yelp scores. *arXiv preprint arXiv:2201.07999*.

Zefang Liu. 2020. Yelp review rating prediction: Machine learning and deep learning models. *arXiv preprint arXiv:2012.06690*.

PartAI. 2024. Dorna llama3 8b instruct.

RadeAI. 2024. Basalam comments and products.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.