Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 3, 2026, 419-431.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Categorization and Aspect Extraction of Online Shopping Comments Using Large Language Models

Mahdi Ahmadlou[1], Abolghasem Daeichian[2,4*], and Ali Reihanian[3,4]

1. Department of Computer Engineering, Sharif University, Tehran, Iran.
2. Department of Electrical Engineering, Faculty of Engineering, Arak University, Arak, Iran.
3. Department of Computer Engineering, Faculty of Engineering, Arak University, Arak, Iran.
4. Research Institute of Artificial Intelligence, Arak University, Arak, Iran.

Article Info

Article History:
Received 11 April 2026
Revised 08 July 2026
Accepted 31 July 2026

DOI:10.22044/jadm.2026.17595.2905

Keywords:
Reviews Categorization, Aspect Extraction, Large Language Models, Sentiment Analysis, Recommender Systems.

**Corresponding author: a-daeichian@araku.ac.ir (A. Daeichian).*

Abstract

The expansion of e-commerce has changed customer purchasing habits, shifting them from brick-and-mortar stores to online venues. In this transition, some fundamental customer behaviors had to change because online shoppers cannot physically feel products and rely heavily on customer reviews for evaluations. However, the lack of structured textual data poses challenges in sifting through numerous, diverse, and sometimes contradictory comments to make informed purchasing decisions. This study proposes a multi-step system for autonomously and intelligently analyzing customer comments to organize comment sections, utilizing general-purpose large language models. First, the proposed system automatically separates comments discussing online shop services from those specifically related to products by tagging them. Then, it extracts various product aspects discussed across all comments. Finally, these comments are categorized based on the extracted aspects. Additionally, a new labeled non-English dataset has been created as a benchmark dataset featuring tagged online-shop-related comments. The experimental results show that the best-performing model was Qwen 2.5, achieving an accuracy of 91.7%.

1. Introduction

In recent years, online shopping has experienced a significant surge in popularity. The COVID-19 pandemic has accelerated the trend of online shopping. A survey on this matter reported that only 11.6% of its participants shopped online in the fall of 2019, while the percentage increased to 51.2% during the pandemic, indicating a substantial rise [1]. Predictions indicate that by 2040, 95% of business transactions will be conducted online [2, 3]. However, online shoppers do not have the opportunity to evaluate product characteristics such as size, texture, and build quality through direct physical inspection, which makes it challenging for them to gauge product quality directly. As a result, online shoppers seek indirect evaluations, including reviews and experiences of other customers.

Research shows that most customers actively write reviews [4, 5], and half of them do so multiple times per month [6]; consequently, the number of reviews is increasing. Reviews and ratings from customers have become the most accessible method for assessing a product's characteristics in online shopping [7].

These reviews provide valuable insights into the product's characteristics and reflect the overall satisfaction of a diverse range of customers, effectively augmenting the in-store experience [8]. Most online shoppers browse reviews before making a purchase. According to a comprehensive study by PowerReviews [6], 95% of customers read reviews during their online shopping journey, with 72% of them doing so multiple times a month. All these findings highlight the critical role that

user reviews play in the customer decision-making process [9]. Nevertheless, research indicates that each customer reads only an average of ten comments before making a purchase, which can lead to inaccurate conclusions. Therefore, e-commerce platforms must display the most relevant and informative comments to potential customers in the most effective way possible.

The comment section is important not only for customers but also for shop owners, who benefit from structured review analysis. It is a space where e-commerce platforms and users can build a connection and mutual trust. A recent survey shows that customers favor online shops that respond promptly to criticism and are twice as likely to purchase from them as from shops that do not respond to comments at all [4]. However, to achieve this, staff of e-commerce platforms need to sift through hundreds or even thousands of comments and differentiate between criticisms of the shop and product-related comments, which is a time-consuming task [10].

A common challenge is the unorganized nature of comments. All parties struggle with the large volume of feedback, and the lack of effective categorization makes it difficult to locate relevant and important feedback. Additionally, each party may require different information from the same comment section. This difficulty in filtering diverse content to find specific details or patterns can complicate the process of drawing meaningful conclusions from the available data [11].

To address these challenges, researchers have increasingly employed sentiment analysis, which automatically identifies opinions and emotions expressed in textual data. Early studies primarily focused on sentiment polarity classification, where an entire review or sentence is categorized as positive, negative, or neutral. While this approach provides a general indication of customer satisfaction, it fails to reveal which specific product features are praised or criticized. Consequently, aspect-based sentiment analysis (ABSA) has emerged as a more fine-grained approach that identifies individual aspects mentioned in a review and determines the sentiment expressed toward each aspect. Despite rapid advances in ABSA, most existing studies have been conducted on English-language datasets, largely due to the abundance of publicly available annotated corpora and linguistic resources. Research on non-English languages remains relatively limited because of the scarcity of labeled datasets and language-specific resources, making the development of accurate ABSA systems for these languages considerably more challenging [14].

Additionally, research on low-resource languages such as Persian has primarily focused on sentiment polarity prediction. While identifying whether an aspect is expressed positively or negatively is valuable, it does not solve the practical problem of organizing large comment sections. To our knowledge, no previous Persian ABSA study has utilized aspect-level sentiment information to automatically categorize customer reviews according to the product aspects they discuss, despite the significant benefits such organization offers for both customers and e-commerce platform owners.

This paper introduces a system to automatically extract aspects from product-related comments in an online shop, considering that it can be employed across different online shops, is language-independent, and utilizes pre-trained models within the proposed framework. This approach ensures that the system does not require significant computational resources and can be easily updated. While existing comment analysis methods often treat reviews monolithically or require extensive task-specific supervised training data (which is especially limiting for non-English, low-resource domains), they fail to provide an autonomous, structured mechanism for breaking down mixed feedback. To fill this gap, the proposed system introduces a multi-stage framework that first separates shop-service-related feedback from product-related feedback, then systematically extracts and categorizes specific product aspects. Furthermore, the system is a fully autonomous, zero-shot framework leveraging general-purpose large language models (LLMs) that operates without manual training data annotations. Another challenge is the lack of a suitable Persian benchmark dataset containing comments and their related aspects. Therefore, a new labeled dataset has been created in this research and is available in [15]. To ensure rigorous evaluation, the benchmark dataset was intentionally constructed to include highly complex real-world text, such as sarcastic remarks and comments with non-standard grammar. The strong performance of the framework across these samples demonstrates its empirical robustness. Although zero-shot LLM outputs are inherently probabilistic, these results suggest that the proposed system can reliably handle conversational noise and implicit sentiments in practical deployment contexts. We conduct extensive experiments to assess the system's performance with unaltered real-world data, focusing on its ability to understand structured comments and identify discussions about shop responsibilities. We also evaluate

aspect identification within comments and investigate the integration of different popular LLMs. The results indicate significant potential for accurately processing user-generated content in complex languages like Persian, demonstrating the system's effectiveness for practical applications.

In brief, this study makes the following contributions:

1. We propose a system for organizing Persian online shopping comments without task-specific training.
2. We introduce a comment categorization pipeline that first distinguishes online-shop-related comments from product-related comments before aspect analysis.
3. We construct a new benchmark dataset containing Persian product reviews annotated with both shop-service labels and generalized aspect labels.
4. We demonstrate that instruction-tuned LLMs effectively transfer semantic knowledge to Persian, enabling competitive aspect extraction without fine-tuning.

The remainder of the paper is outlined as follows: Section 2 reviews the literature on aspect-based sentiment analysis methods and discusses their advantages and disadvantages. In Section 3, the design of the proposed system is explained in detail. The results, including comparison results, are presented and discussed in Section 4. Finally, in Section 5, the paper is concluded, and the challenges of the proposed system, along with suggestions for improvements, are discussed.

2. Literature Review

The global field of sentiment analysis began in the early 2000s, with the first works focusing on polarity detection at the document and sentence levels [12, 13]. A significant change occurred around 2010 with the first uses of the term aspect-based sentiment analysis (ABSA). ABSA involves extracting key information from written text and is a more fine-grained type of sentiment analysis [14]. While standard sentiment analysis identifies the overall feeling of a text, ABSA dives deeper. It pinpoints specific features or "aspects" within the text and determines the sentiment expressed towards each one individually. Early methods relied on traditional machine learning, but by the mid-2010s, deep learning began to predominate due to the availability of more benchmark datasets [15, 16]. Some research has implemented recurrent neural networks (RNNs) to capture long-term dependencies [17]. , though they suffer from sequential bottlenecks and high training times. Meanwhile, convolutional neural networks effectively extract local n-gram features [18] but struggle with long-range context. Subsequently, transformer models like BERT [19] resolved context issues but require extensive task-specific fine-tuning and annotated data.

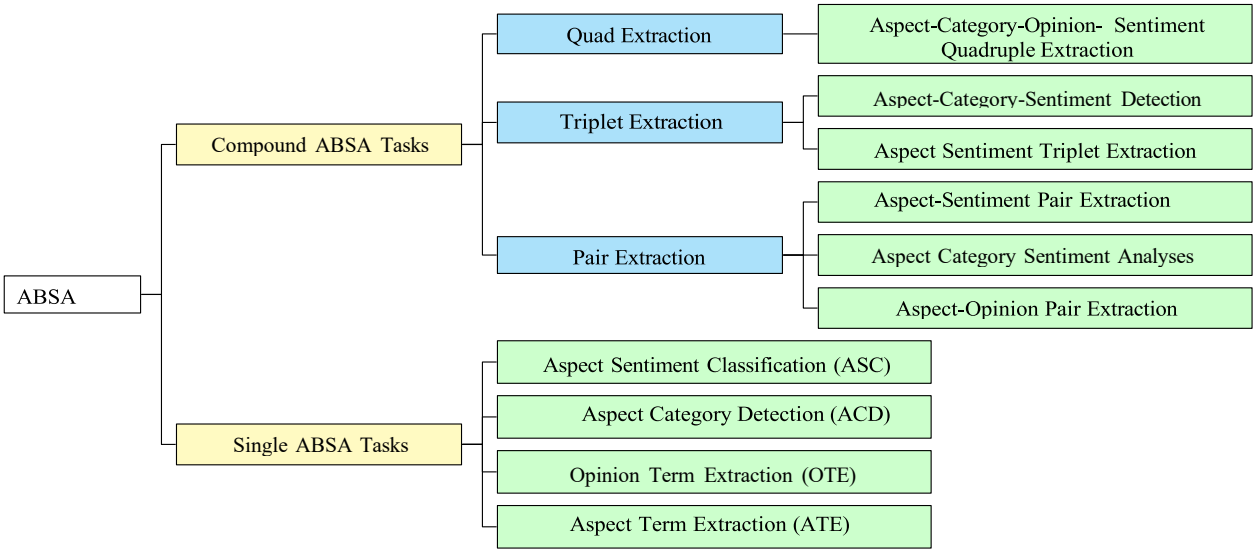

Figure 1. Taxonomy and relationships of Aspect-Based Sentiment Analysis (ABSA) tasks, illustrating the hierarchical division between single tasks and compound tasks.

Aspect-based sentiment analysis (ABSA) tasks can be organized into two categories of single and compound, each addressing different facets of sentiment analysis [20] (see Figure 1). Single ABSA tasks focus on extracting a specific component within a text. For instance, the Aspect Term Extraction (ATE) task focuses on identifying a key term related to a specific aspect (such as a product attribute), while Opinion Term Extraction (OTE) seeks to extract a term expressing an opinion. Aspect Category Detection (ACD) task is related to classifying an opinion term into a

predefined category, and Aspect Sentiment Classification (ASC) aims to determine the sentiment (positive, negative or neutral) associated with an aspect. For example, consider the sentence "The battery life is excellent." as a comment. Considering single ABSA tasks, the term "battery life" can be identified as an aspect, it can be classified, and the sentiment towards this aspect can be determined as positive. Compound ABSA tasks combine several individual ABSA tasks which are shown in Figure 1.

The emergence of transformer-based architectures, in particular LLMs, has driven a paradigm shift in natural language understanding and sentiment analysis. For example, BERT [19] introduced deep, bidirectional contextualized representations pretrained on large unmarked datasets and then fine-tuned on the target task. Current instruction-tuned models such as FLAN-T5 and GPT-3 exhibit high performance even under few-shot or zero-shot settings without requiring any additional parameter updates [21]. Recently, a paradigm shift occurred as Generative Pre-trained Transformer (GPT) models introduced instruction-following generative pipelines for ABSA. Rather than relying solely on traditional token-classification architectures, these models can perform joint aspect extraction and categorization through prompt design. It was found in [22] that practical prompt engineering allows GPT-4 to outperform BERT models in few-shot settings.

However, unlike the global field, the aspect-based sentiment analysis in Persian was first utilized in deep learning around 2018-2019, primarily due to challenges such as the limited availability of labeled datasets and the language's complexity. Some studies have achieved an acceptable accuracy on a new product dataset using that approach [23-26]. Following global trends, researchers adopted transformer-based models, leading to the creation of ParsBERT, a monolingual BERT model trained on a large Persian corpus, which outperformed multilingual models [27, 28]. Other BERT-based approaches like [29] also exist and recently some studies used GPT for text classification [30]. It is worth mentioning that the Pars-ABSA and FABSA datasets, compiled from reviews on the Iranian platforms, became a key benchmark [31] There is a limited number of datasets for Persian ABSA, and none of them have online-shop-related comments labeled to separate them from other comments, prompting us to develop a new dataset. The application of GPT-based generative models to the Persian ABSA task is largely unexplored. Initial Persian research shifted focus to BERT architectures due to the unavailability of powerful models like GPT-3. Recent studies have primarily focused on coarse-grained sentiment analysis. Recent papers highlight that GPT models on Persian ABSA have been overlooked, underscoring a research gap [32]. To mitigate data scarcity, recent studies like [32] introduced cross-lingual deep learning frameworks to transfer models from high-resource languages (like English) to Persian using cross-lingual embeddings. While effective at bridging the resource gap, these traditional approaches typically require task-specific fine-tuned classifiers and annotated parallel data, limiting their adaptability to novel, unobserved domains or zero-shot conversational styles.

This study proposes a multi-step framework using Large Language Models for review processing, focusing on Aspect Term Extraction (ATE) and Aspect Category Detection (ACD) in online shopping comments. By structuring unstructured user feedback into targeted aspect terms and predefined categories, the framework enables efficient opinion organization without requiring full sentiment polarity classification. In this study, an end-to-end, zero-shot framework utilizing freely available multilingual LLMs specifically for hierarchical aspect term extraction and aspect category detection in Persian online shopping reviews are presented. We also introduce a novel hierarchical aspect tagging system that goes beyond the standard ABSA task in e-commerce. While previous work, like the Pars-ABSA dataset, extracts a flat list of product-related aspects (e.g., "camera," "quality"), it lacks direct utility for business intelligence. Our system tags aspects that are business-relevant, such as Shipping and Logistics, and Customer Service, and extracts aspects related to the product, such as Price and Value. Recognizing the lack of comprehensive datasets in this area, especially the version with tagged online shop related comments, we also developed a new database to support other researchers in this field. We discuss the characteristics of this dataset in detail in the dataset section. This study can transform the standard comment section into a more organized and usable area, enabling online retailers to directly link customer feedback to operational departments and pinpoint structural strengths and weaknesses.

3. Proposed Method

This paper addresses three main objectives:

1) Identifying product aspects discussed in reviews
2) How can the different product aspects expressed in product reviews be categorized?

3) How can the answers to the first two research questions be utilized to benefit all parties involved in online shopping, including customers, online shops, and manufacturers?

To this end, the primary focus of this paper is on extracting aspect terms and identifying aspect categories, which incorporate two primary tasks of ATE and ACD (see Figure 2, as an example). An LLM is employed in the process of implementing the proposed system, which plays an important role in the two mentioned primary tasks.

We chose to use the newly released LLM by Meta, named Llama3. Llama3 is an open-source model, which can be freely downloaded.The precise adjustment of parameters, which greatly affects both the training speed and the output quality, is vital for an LLM. The adjustment of parameters of an LLM should be done in a way that strikes a good balance between its efficiency and performance. The parameter settings of the LLM which is used in the proposed system are outlined in Table 1.

Figure 3 illustrates the proposed system architecture. Additionally, the consolidated pipeline workflow is summarized step-by-step in Algorithm 1.

Table 1. The parameters of the LLM in the proposed method.

Parameter	Value	Description
N-Ctx	5900	Maximum number of tokens
Temp	0	Adjusts randomness and diversity of generated predictions
Seed	42	Controls randomness, ensuring reproducible results
Prompting strategy		Zero-shot
Precision		FP16 (16-bit)
Top-p		1.0
Top-k		40

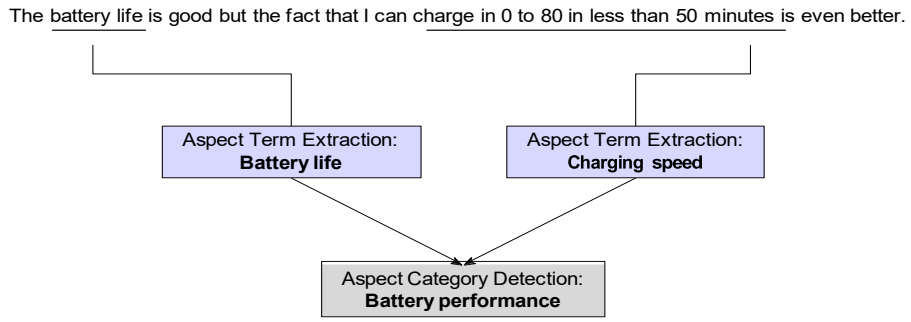

Figure 2. An example of Aspect Term Extraction (ATE) and Aspect Category Detection (ACD).

Algorithm 1. Pseudocode of the proposed pipeline.

Input:
C: Customer comments
S: Shop service descriptions
M: Instruction-tuned LLM
Output:
D: Aspect-annotated dataset

#	Procedure
1	Collect comments C from online shopping websites.
2	for each comment $c_i \in C$ do
3	shop_tag ← LLM (c_i, Prompt_shop, S)
4	aspects_specific←LLM (c_i, Prompt_extract)
5	end for
6	Aggregate all extracted aspects.
7	Merge synonymous aspects using LLM.
8	Select generalized aspect list A.
9	for each comment $c_i \in C$ do
10	for each aspect $a_j \in A$ do
11	Label (c_i, a_j) ← LLM (c_i, Prompt_classify, a_j)
12	end for
13	end for
14	Human annotators independently verify generated labels.
15	Retain samples with unanimous agreement.
16	Evaluate predictions using Accuracy, Precision, Recall, and F1-score.
17	Return D.

While a single monolithic prompt could theoretically handle extraction, decomposing the workflow into sequential modular stages (service filtration followed by aspect consolidation) is intentionally designed to isolate domain noise, prevent LLM hallucinations in low-resource settings, and manage context window constraints. The schematic encompasses the initial stages of raw comment acquisition from existing online shops to the final analysis of the system's output. The results are then compared against a benchmark dataset that was independently annotated by a three-person team. It is worth noting that, although the multi-step pipeline achieves high accuracy, it incurs a cumulative computational and latency overhead due to multiple sequential API/inference calls per comment. All prompts were engineered following standard zero-shot instruction principles. A detailed explanation of all stages is provided below:

Step 1: Customers submit comments after purchasing products from online shopping websites. These comments contain information about product characteristics as well as online-shop services (e.g., delivery, packaging, or customer support). Such comments provide valuable

feedback for customers, online shops, and manufacturers, supporting purchasing decisions, service improvements, and product quality assessments.

Step 2: Customer comments are collected either directly from online shop owners or via web crawling, where permitted. The dataset includes reviews of multiple product categories—such as guitars, laptops, and mobile phones—collected from several Iranian e-commerce websites to ensure diversity in products and user opinions.

Step 3: A prompt containing the online shop's name, services, and responsibilities is provided to the LLM. For each comment, the LLM determines whether it discusses any shop-related service or responsibility and returns a binary response ("yes" or "no"). This step separates shop-related comments from product-related comments. The prompt is presented in Table 2.

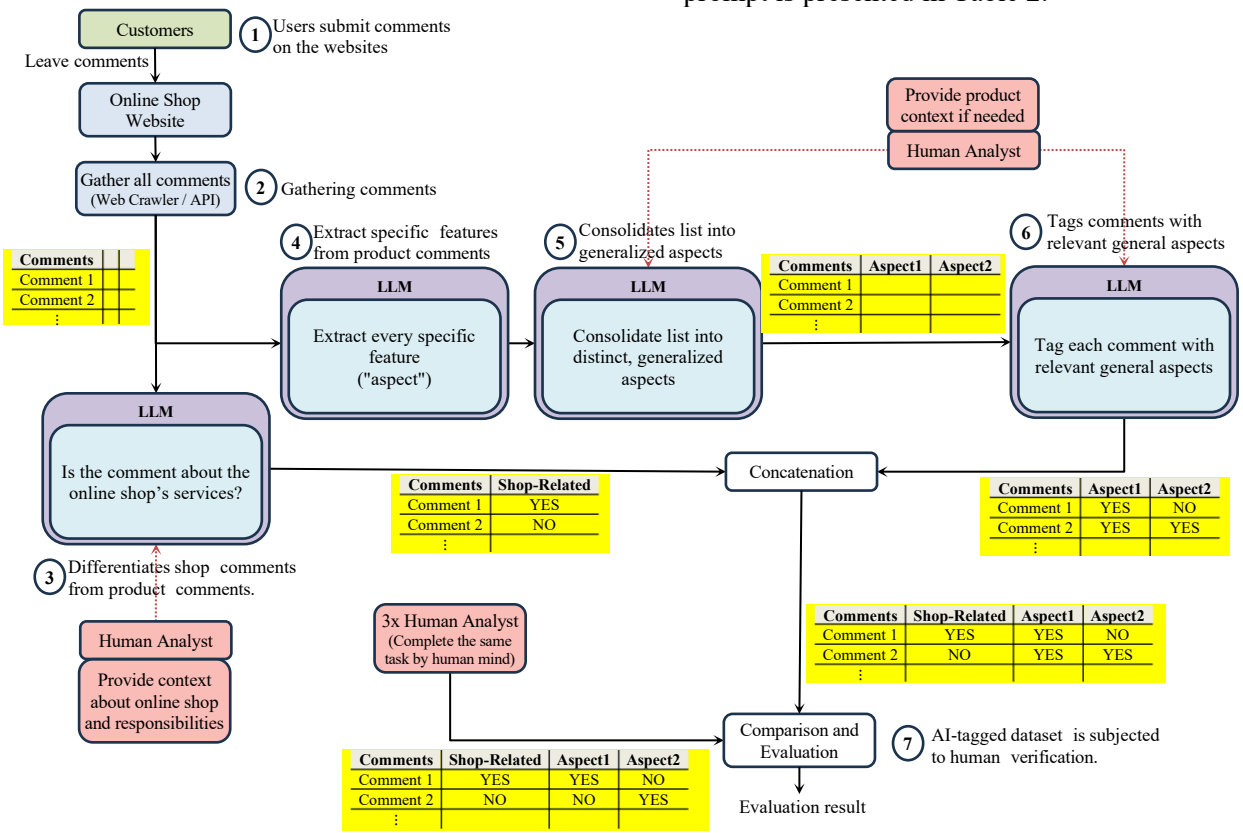

Figure 3. The graphical diagram of the proposed system.

Table 2. The prompts related to the steps of the proposed system.

Step	Task	Prompt
3	Determining the online shop-related comments	Comment: {comment} I provide you with a comment about a {product} on {website_name}. The services of {website_name} are {services}, and its responsibilities are {responsibility}. Determine if the comment is related to {website_name}'s services or responsibilities. Respond with one word only: (Yes) or (No).
4	Aspect extraction	I'm giving you a comment from a buyer. I want you to identify the most general and different aspects mentioned about the product in the comment for aspect-based sentiment analysis. Make sure those aspects are not about the e-commerce site, return policy, or delivery. Please provide the aspects only, without specific details or further explanation. The aspects should be short and precise. The aspects should not be about the online shop and should only be about the product. The comment is as follows: Comment: \{comment\}
5	Categorizing the extracted aspects	Categorize the list I give you and give me the most repeated categories. The list is: \{list of aspects\}
6	Finding the categorized aspects in each comment	Comment: \{comment\} I have a comment from my online shop post, along with an aspect. Please determine if this comment mentions anything about \{aspect\} of the product. Respond with one word either (Yes) or (No) without any explanation.

Step 4: Each product-related comment is submitted to the LLM with a prompt requesting all aspects mentioned in the comment. The extracted aspects are stored for each product, and the process is repeated for all products in the dataset. The corresponding prompt is shown in Table 2. To investigate the effect of prompt design, two prompt configurations are evaluated. The first includes additional product information (brand and category), while the second uses only the comment text. Their performance is compared in the

experimental section to determine whether the additional context improves aspect extraction.

Step 5: The extracted aspects of each product are consolidated into a set of generalized aspects. Since semantically similar aspects (e.g., battery and battery life) are often extracted separately in Step 4, the 15 most frequent aspects are first selected and provided to the LLM. The model is then instructed to merge semantically related aspects and produce a list of generalized aspects. Restricting input to the most frequent dataset-derived aspects constrains the model to semantic organization rather than free generation, reducing hallucinations and subjective bias. The corresponding prompt is presented in Table 2.

Step 6: For each product-related comment, a prompt containing the comment and the generalized aspect list obtained in Step 5 is submitted to the LLM. The model identifies all aspects from the predefined list mentioned in the comment. Repeating this process for every comment produces the final aspect-annotated dataset, linking comments with the generalized aspects of 19 products. The prompt is shown in Table 2.

Step 7: In this step, the dataset of comments generated previously, along with their associated aspects, was independently distributed to three annotators who were asked whether a specific aspect is discussed in the related comment or not. The annotators were blinded to the AI-generated outputs from Step 6 and had no access to the aspect assignments produced by the LLM. For example, the comment “it has good quality but terrible camera” was sent to them, and they were asked if any of the aspects “quality”, “camera”, ”sound”, and ”design” is discussed in the comment or not. We followed a Double-Blind, Independent Annotation Protocol with a Majority-Vote Consensus (2/3 rule). To guarantee objective annotations, the three human annotators worked completely independently without any communication during the annotation process and were fully blinded to both each other's inputs and the model predictions. The human majority labels were used as the gold-standard benchmark to evaluate the model's independent binary predictions, enabling an unbiased evaluation of Precision, Recall, and F1-score. The Llama model, as well as several other prominent LLMs, is used in Steps 4, 5, and 6, and the performance is evaluated which can be viewed in Section 4.

4. Experimental Results

In this section, we present experimental results, providing a comprehensive overview of our findings. We begin by detailing the dataset created in this research, offering insights into its composition and significance. Subsequently, we evaluate the performance of the proposed system from various perspectives, highlighting its strengths and areas for improvement. Finally, the performance of the proposed system, when considering some different popular LLMs within its framework, is investigated.

4.1. Dataset

One of the main challenges in performing natural language processing tasks for non-English documents is the shortage of high-quality datasets in the related language. For example, one of the primary goals of this paper is to analyze comments related to different products on various online shops and identify the aspects of the products discussed in each comment. As previously mentioned, the paper aims to do this for comments written in Persian. However, no standard benchmark dataset exists in which comments are in Persian and their related aspects are available. One possible reason for this shortcoming is that Persian utilizes a rich system of morphology and syntax, which can lead to a high degree of ambiguity in natural language processing tasks. The language features complex verb conjugations, diverse affixes, and a significant use of idiomatic expressions, which are difficult to interpret correctly without extensive understanding of cultural and contextual nuances. Additionally, performing text preprocessing tasks such as tokenization and stemming for Persian documents is complicated, especially because the language's writing system is right-to-left and has a cursive style. The lack of large labeled datasets in Persian for training machine learning models further exacerbates the difficulty of developing accurate automatic aspect identification tools for comments in Persian, making it a particularly challenging endeavor in the field of computational linguistics.

In order to overcome the problem explained above, a new labeled dataset is created in this research, in which the comments about different products (related to an Iranian online shop) along with their related aspects are available. Naturally, all comments in this dataset are written in Persian. The dataset has some unique features. For example, the comments, in which online shop responsibilities or services were discussed, are separated from those specifically related to the products, by the related tags. This dataset facilitates the training and fine-tuning of different machine learning models, designed for performing different natural language processing tasks considering Persian language,

with the aim of enhancing their accuracy and effectiveness. On the other hand, the dataset can be considered a benchmark for other researchers to test and evaluate the performance of their proposed systems in analyzing Persian comments.

To construct such a dataset, in this research, initially, a web crawler was used to collect the comments from the related online shop websites. To evaluate the system's performance objectively and avoid any automation or confirmation bias, an independent, double-blind annotation process was conducted to establish the ground-truth baseline. The 1338 raw comments were distributed to three human annotators along with a blanket list of candidates generalized aspects for that product category (e.g., *quality, camera, sound, design*), without any labels. For each candidate aspect, the annotators answered a binary question: *"Does this comment discuss this specific aspect?"* Crucially, the annotators were completely blinded to the AI-generated outputs from Step 6 and had no knowledge of whether the LLM had predicted a "Yes" or "No" for any given aspect. To ensure human annotation quality, Inter-Annotator Agreement (IAA) was computed across all 1338 annotated comments, yielding a Fleiss' Kappa of $\kappa = 0.671$ (Substantial Agreement). To create a high-fidelity ground-truth evaluation set, we retained only the 998 comments that achieved unanimous consensus (3/3 agreement), excluding the 390 comments with split decisions Finally, the human consensus labels were used as the gold-standard benchmark to evaluate the model's independent binary classifications, allowing for an unbiased assessment of Precision, Recall, and F1-score. Several comments included in the final dataset contain defects, such as grammatical or phonetic issues, elements of sarcasm, ambiguity and implicit references to aspects. Since the goal of the research was to test the proposed system under real conditions, no corrections were made to the mentioned comments. Analyzing such comments is more challenging compared to working with pre-processed cleaned ones. By preserving the original comments in the final dataset, it can be accurately assessed how well the proposed system can handle the unaltered real-world data while considering their challenges and complexities. Also, the acceptable performance of the proposed system in processing such complicated data can be considered as a reliable indicator of the system's performance in practical applications.

Additional characteristics of the final dataset are presented in Tables 3, 4, and 5. Specifically, Table 3 shows the distribution of comments across product categories, Table 4 summarizes the annotated aspects for each category, and Table 5 reports descriptive statistics of the dataset, including the total number of comments, total number of words, and average comment length.

Table 3. Approved and Rejected comments counts for 19 products.

Product Name	Approved	Rejected
Samsung A54	110	77
Beats Headphone	29	21
Canon Camera	21	11
Tent	40	7
Zigma Espresso Machine	181	56
Guitar	29	5
Vacuum Cleaner	44	2
Shoe	31	9
Keyboard	19	11
Backpack	25	21
Washing Machine	37	13
Refrigerator	41	21
Samsung S24	39	11
Sunscreen	32	17
Watch	92	6
Pressure Cooker	106	17
Portable Modem	59	21
Power Bank	29	13
Printer	24	11

Table 4. Product Categories and Corresponding Annotation Aspects.

Product Category	Annotation Schema (Evaluated Aspects)
A54 (Smartphone)	Phone Quality, Price and Value, Phone Hardware, Phone Performance, Packaging and Delivery
Backpack	Quality and Performance, Physical Characteristics, Practicality
Beats (Headphones)	Sound-related aspects, Comfort and ergonomics aspects, Build and design aspects, Battery and power aspects
Canon (Camera)	Product Quality, Beginner-Friendliness, Price and Value, Image Quality
Espresso Machine	Quality-related aspects, Price and value, Performance and functionality, Design and features, Ease of use and usability
pressure Cooker	Quality, Performance, Design, Price, Ease of use
Guitar	Quality, Packaging, Price, Accessories
Keyboard	Quality/Product quality, Price, Value, Performance
Portable Modem	Quality, Battery, Performance, Price/Value, Design/Physical
Power-bank	Quality-related, Charging-related, Battery-related, Design and Build, Performance-related
Printer	Quality-related aspects, Cartridge and Ink aspects, Cost and Price aspects
Refrigerator	Price, Quality, Installation, Cooling, Size
S24 (Smartphone)	Quality, Design, Performance, Camera, Price
Shoes	Quality, Size, Price, Comfort, Material
Sunscreen	Physical Properties, Economic and Practical, Performance, Ingredients
Tent	Quality, Size, Material, Price, Design
Vacuum Cleaner	Performance, Design and Appearance, Quality and Durability, Practicality and Convenience, Power and Battery
Washing Machine	Quality, Size, Performance, Price
Watch	Quality, Price, Appearance, Performance, Comfort, Water resistance

Table 5. Some characteristics of the final Dataset.

Data	Numbers
Words per aspects	17.95
Total No. of Words in Dataset	39401
Total No. of Aspects in Comments	76
Average No. of Aspects in Comments	2.22
Total Number of Products	19
Total Number of Shops	19
Total Number of comments	998
Average No. of Words in each Comment	33
Vocabulary Size (Unique Words)	4943 Comments per Product
Most Frequent Aspect Category	price
Overall Total Positive Responses	2,193
Overall Total Negative Responses	2,572

4.2. Results

In this sub-section, the performance of the proposed system is analyzed from different points of view. Thus, first, it is examined whether the proposed system is capable of understanding the purpose of the comments with a special kind of structure. Then, the performance of the proposed system in identifying the comments, in which online shop responsibilities or services were discussed, is evaluated. Finally, the performance of the proposed system in properly identifying the aspects of the product-related comments is analyzed.

4.2.1. The capability to understand comments with complex language structures

The proposed system is capable of understanding the purpose of comments with complex negative structures, as illustrated in Table 6-Example 1. Additionally, it demonstrated good performance in understanding comments in which the presence of a negative verb alters the semantic interpretation of a sentence, as shown in Table 6-Example 2. Thus, considering this example, we can conclude that the proposed system has a notable ability to understand the complex linguistic structures of the Persian language.

Table 6. The capability of the proposed system in understanding complex linguistic structure of Persian language.

Example 1	
Sentence	تازه خریدم نمیشه در مورد کیفیت نظر داد.
Meaning	I just bought it, so I can't comment on the quality yet.
Result	The proposed system understood that the quality was not discussed because of a negative verb.
Example 2	
Sentence	گوشی خوبی بود ولی نتونستم در مورد گارانتیش نظر بد ندم.
Meaning	It was a good phone, but I couldn't help leaving a negative comment about its warranty.
Result	The proposed system understood that the warranty was discussed in a situation where a negative verb was used.

4.2.2. Identifying online shop related comments

In this part, the performance of the system in identifying comments discussing online shop responsibilities or services is evaluated (for more information, please refer to Step 3 of the proposed system). For this reason, the four metrics of Accuracy, Precision, Recall, and F1-Score were calculated:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where *TP* (True Positives) is the number of instances the model correctly predicted as positive, *TN* (True Negatives) is the number of instances the model correctly predicted as negative, *FP* (False Positives) is the number of instances the model incorrectly predicted as positive, and *FN* (False Negatives) is the number of instances the model incorrectly predicted as negative.

As illustrated in Figure 4, the model achieved high classification performance in the mentioned task by reaching a Precision of 97.04%, Recall of 93.28%, F1-score of 94.91%, and Accuracy of 98.18%. Thus, it can be concluded that the proposed system is capable of reliably distinguishing comments discussing online shop responsibilities or services from those specifically related to the products.

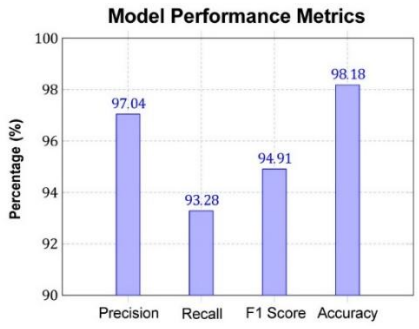

Figure 4. The performance of the system on identifying online shop or product related comments.

4.2.3. Identifying the aspects of the product-related comments

The performance of truly identifying the aspects in product-related comments is evaluated in two scenarios, which are explained in Step 4 of the proposed system (for more information, see Section 3). In the first scenario, the aspect

extraction prompt was accompanied by additional information about the product to enhance the system's performance. It should be noted that this extra information also made the prompt more complex and labor-intensive, requiring human intervention to input the details. This information concerned the brand and the category of the related product. In the second scenario, no extra information was included in the aspect extraction prompt of the system.

The impact of considering these two scenarios on the aspect identification performance of the proposed system in product-related comments is presented in Table 7. The results demonstrate that appending extra information about the product to the prompt mostly led to worse results compared to the scenario where no additional information was included. Even in the Llama 3.3 model, where providing additional information to the model improved performance, the differences were not statistically significant, as indicated by a z-score of 1.21 and a p-value of approximately 0.2261.

By evaluating the performance of the proposed system, it can be concluded that the system has a considerable potential in reliably and accurately processing user-generated content for e-commerce applications and other related domains.

Table 7. Details of the aspect identification performance of the proposed system in product-related comments, from the two scenarios, separately listed for each 19 products of the final dataset.

Dataset	No extra info was considered in the related prompt				Extra info was considered in the related prompt			
	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy
Pressure Cooker	94.91	75.09	83.84	85.09	96.30	75.09	84.38	85.47
Watch	89.92	78.11	83.60	83.51	91.09	83.33	87.04	87.32
Sunscreen	98.70	80.00	88.37	84.38	98.70	80.00	88.37	84.38
Samsung S24	89.61	80.23	84.66	87.18	87.01	79.76	83.23	86.15
Refrigerator	94.12	84.21	88.89	90.24	96.47	84.54	90.11	91.22
Printer	98.04	87.72	92.59	88.89	98.04	89.29	93.46	90.28
Power Bank	100.0	73.58	84.78	80.69	97.44	69.72	81.28	75.86
Portable Modem	96.58	86.50	91.26	90.85	94.52	85.19	89.61	89.15
Washing Machine	96.61	75.00	84.44	85.81	98.31	77.33	86.57	87.84
Backpack	96.08	77.78	85.96	78.67	94.12	78.69	85.71	78.67
Keyboard	97.22	76.09	85.37	84.21	86.11	79.49	82.67	82.89
Shoe	93.69	71.23	80.93	77.73	95.65	81.48	88.00	88.39
Vacuum Cleaner	82.88	85.98	84.40	84.55	98.31	77.33	86.57	87.84
Guitar	98.28	87.69	92.68	92.24	87.93	96.23	91.89	92.24
Zigma Espresso Machine	95.75	68.70	80.00	77.59	96.23	71.20	81.85	80.00
Tent	88.46	77.53	82.63	85.07	94.87	77.89	85.55	87.50
Canon Camera	85.71	80.00	82.76	85.14	85.71	85.71	85.71	85.71
Beats Headphone	92.19	83.10	87.41	85.93	95.31	85.92	90.37	88.79
Samsung A54	93.43	69.34	79.6	92.96	92.16	71.48	80.82	82.91
Average	**93.80**	**78.84**	**85.48**	**84.30**	**93.91**	**80.51**	**86.48**	**85.93**

4.2.4. Considering other well-known LLMs in the framework of the proposed system

Among the evaluated LLMs, Llama 3.3 and Qwen 2.5 achieved higher Precision values compared with the other evaluated scenarios, demonstrating their stronger capability to accurately identify relevant aspects from product-related comments. However, the high Precision achieved by Llama 3.3 was accompanied by a relatively lower Recall, which limited its overall performance and resulted in lower final Accuracy. In contrast, Qwen 2.5 provided a more balanced performance by maintaining comparable Precision while achieving substantially higher Recall (87.8% compared with 78.8% for Llama 3.3), leading to superior overall performance (Accuracy of 91.7%). Therefore, among the evaluated models without additional product information, Qwen 2.5 demonstrated the most effective balance between correctly identifying relevant aspects and minimizing missed aspects. A comprehensive comparison of all evaluation metrics across the different evaluated LLMs is presented in Figure 5. In addition, the statistical significance analyses for different LLMs are given in Table 8.

Table 8. Pairwise model performance comparisons with z-scores and corresponding p-values.

Model Comparison	Z-Statistic	p-value	Significance (α=0.0167)
Llama 3.3 vs. Qwen 2.5	6.28	3.30×10^{-10}	True
Llama 3.3 vs. Mistral Large	9.0	2.289×10^{-19}	True
Qwen 2.5 vs. Mistral Large	2.76	0.00586	True

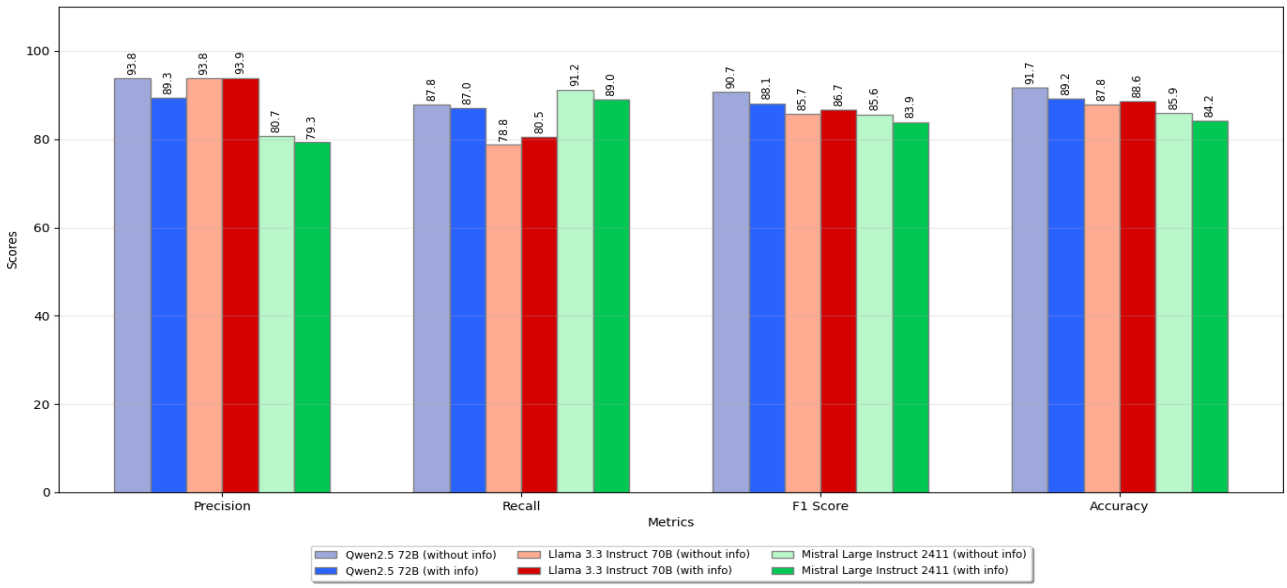

Figure 5. Results of considering some different popular LLMs in the framework of the proposed system.

5. Conclusion

In this paper, a system has been introduced with the purpose of automatic extraction of aspects from product-related comments of an online shop. In order to achieve the mentioned goal, the proposed system employs a large language model (LLM). In addition to aspect extraction, the proposed system is capable of separating the comments, in which online shop responsibilities or services were discussed, from those specifically related to the products, in the early stages of the system, by tagging them. Also, the proposed system can analyze the comments that are in Persian (which is a complex natural language), even in conditions such that they include grammatical or phonetic issues, elements of sarcasm, ambiguity and implicit references to aspects. Furthermore, to address the challenge of the lack of a suitable benchmark dataset, in which the comments are in Persian and their related aspects are available, a new labeled dataset was created in this research, as the output of the proposed system, in which the comments (in Persian) about different products (related to an Iranian online shop) along with their related aspects are available.

Extensive experiments were conducted in this research in order to evaluate the performance of the proposed system in handling unaltered real-world data while considering their challenges and complexities. First, it was examined whether the proposed system is capable of understanding the purpose of the comments with special kind of structures. Then, the performance of the proposed system in identifying the comments, in which online shop responsibilities or services were discussed, was evaluated. After that, the performance of the proposed system in properly identifying the aspects discussed in each comment was analyzed. Finally, the performance of the proposed system, when some different popular LLMs were considered in its framework, was investigated. The experimental results indicated the considerable potential of the proposed system in reliably and accurately processing user-generated content in complex natural language (such as Persian), which can be considered as a reliable indicator of the system's performance in practical applications.

Data Availability Statement

The data that support the findings of this study are openly available on Github at: https://github.com/mehtix1/PABSA-GPT.

References

[1] M. Young, J. Soza-Parra, and G. Circella, "The increase in online shopping during COVID-19: Who is responsible, will it last, and what does it mean for cities?," *Regional Science Policy & Practice,* vol. 14, pp. 162–178, 2022.

[2] B. Rohrßen, "Digital Distribution: Online Sales and Online Platforms," in *VBER 2022: EU Competition Law for Vertical Agreements: Digital, Dual, Exclusive and Selective Distribution plus Franchising*: Springer, 2023, pp. 167–175.

[3] D. Holubinka, V. Vysotska, S. Vladov, Y. Ushenko, M. Talakh, and Y. Tomka, "Intelligent system for recognizing tone and categorizing text in media news at an electronic business based on sentiment and sarcasm analysis," *International Journal of Information Engineering and Electronic Business (IJIEEB),* vol. 17, no. 1, pp. 90–139, 2025.

[4] Bright-Local, "Local consumer review survey 2024," 2024. [Online]. Available: https://www.brightlocal.com/research/local-consumer-review-survey/

[5] N. Arizal, Nofrizal, W. Dwika Listihana, and Hadiyati, "Gen z customer loyalty in online shopping: an integrated model of trust, website design, and security," *Journal of Internet Commerce,* vol. 23, no. 2, pp. 121–143, 2024.

[6] Powerreviews, "Survey: The Ever-Growing Power of Reviews 2024," *powerreviews*, 2024. [Online]. Available: https://www.powerreviews.com/power-of-reviews-2023/

[7] B. Ganguly, P. Sengupta, and B. Biswas, "What are the significant determinants of helpfulness of online review? An exploration across product-types," *Journal of Retailing and Consumer Services,* vol. 78, p. 103748, 2024.

[8] Z. Wang, S.-j. Hu, S.-f. Niu, S.-y. Li, W.-d. Liu, and L.-y. Huang, "Research on product design improvement method based on online review and improvement importance performance competitor analysis," *Expert Systems with Applications,* vol. 279, p. 127400, 2025.

[9] J. Yi and Y. K. Oh, "The informational value of multi-attribute online consumer reviews: A text mining approach," *Journal of Retailing and Consumer Services,* vol. 65, p. 102519, 2022.

[10] D. Dalli, D. D'Acunto, and A. Tuan, "How online reviewers and actual customers evaluate their shopping experiences: evidence from an international retail chain," *Mercati e competitività: 3,* pp. 163–180, 2018.

[11] O. Coban, M. Yağanoğlu, and F. Bozkurt, "Domain effect investigation for BERT models fine-tuned on different text categorization tasks," *Arabian Journal for Science and Engineering,* vol. 49, no. 3, pp. 3685–3702, 2024.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002),* pp. 79–86, 2002.

[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004.

[14] Y. Liu, J. Shi, F. Huang, J. Hou, and C. Zhang, "Unveiling consumer preferences in automotive reviews through aspect-based opinion generation," *Journal of Retailing and Consumer Services,* vol. 77, p. 103605, 2024.

[15] A. P. Richard Socher, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *the Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631-1642, 2013.

[16] J. Feng, S. Cai, and X. Ma, "Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm," *Cluster Computing,* vol. 22, no. Suppl 3, pp. 5839–5857, 2019.

[17] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.

[18] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems,* vol. 108, pp. 42–49, 2016.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, vol. 1, pp. 4171–4186, 2019.

[20] D. Kirange, R. R. Deshmukh, and M. Kirange, "Aspect based sentiment analysis semeval-2014 task 4," *Asian Journal of Computer Science and Information Technology (AJCSIT) Vol,* vol. 4, 2014.

[21] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines,* vol. 30, pp. 681–694, 2020.

[22] P. F. S. a. P. Huoviala, "Large language models for aspect-based sentiment analysis," *arXiv preprint,* arXiv:2310.18025, 2023.

[23] S. M. Al-Ghuribi, S. A. M. Noah, and S. Tiun, "Unsupervised semantic approach of aspect-based sentiment analysis for large-scale user reviews," *IEEE Access,* vol. 8, 2020.

[24] A. N. Karimvand, R. S. Chegeni, M. E. Basiri, and S. Nemati, "Sentiment analysis of persian instagram post: a multimodal deep learning approach," in *2021 7th International Conference on Web Research (ICWR)*, IEEE, pp. 137–141, 2021.

[25] S. Shumaly, M. Yazdinejad, and Y. Guo, "Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings," *PeerJ Computer Science,* vol. 7, p. e422, 2021.

[26] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment analysis of persian movie reviews using deep learning," *Entropy,* vol. 23, no. 5, p. 596, 2021.

[27] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," *Neural Processing Letters,* vol. 53, pp. 3831–3847, 2021.

[28] Ghasemi, Ali Reza, and Javad Salimi Sartakhti. "Multilingual Language Models in Persian NLP Tasks: A Performance Comparison of Fine-Tuning Techniques." *Journal of AI and Data Mining,* vol. 13, no.1, pp. 107-117, 2025.

[29] H. Jafarian, A. H. Taghavi, A. Javaheri, and R. Rawassizadeh, "Exploiting BERT to improve aspect-based sentiment analysis performance on Persian language," in *2021 7th International Conference on Web Research (ICWR)*, IEEE, pp. 5–8, 2021.

[30] T. S. Ataei, K. Darvishi, S. Javdan, A. Pourdabiri, B. Minaei-Bidgoli, and M. T. Pilehvar, "Pars-off: a benchmark for offensive language detection on farsi social media," *IEEE Transactions on Affective Computing,* vol. 14, no. 4, pp. 2787–2795, 2022.

[31] G. Kontonatsios *et al.*, "FABSA: An aspect-based sentiment analysis dataset of user reviews," *Neurocomputing,* vol. 562, p. 126867, 2023.

[32] A. Abaskohi *et al.*, "Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 2189–2203, 2024.

دسته‌بندی و استخراج جنبه‌های نظرات خرید آنلاین با بهره‌گیری از مدل‌های زبانی بزرگ

مهدی احمدلو[۱]، ابوالقاسم دائی‌چیان[۲،۴*] و علی ریحانیان[۳،۴]

[۱] دانشکده مهندسی کامپیوتر، دانشگاه شریف، تهران، ایران.

[۲] گروه مهندسی برق، دانشکده فنی و مهندسی، دانشگاه اراک، اراک، ایران.

[۳] گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اراک، اراک، ایران.

[۴] پژوهشکده هوش مصنوعی، دانشگاه اراک، اراک، ایران.

ارسال ۲۰۲۶/۰۴/۱۱؛ بازنگری ۲۰۲۶/۰۷/۰۸؛ پذیرش ۲۰۲۶/۰۷/۳۱

چکیده:

گسترش تجارت الکترونیکی، عادات خرید مشتریان را دگرگون کرده و آنان را از فروشگاه‌های فیزیکی به بسترهای آنلاین سوق داده است. در این گذار، برخی رفتارهای اساسی مشتریان نیز ناگزیر تغییر کرده‌اند؛ زیرا خریداران آنلاین نمی‌توانند محصولات را از نزدیک لمس کنند و برای ارزیابی آن‌ها تا حد زیادی به نظرات سایر مشتریان متکی هستند. بااین‌حال، نبود داده‌های متنی ساختاریافته از نظرات خریداران، بررسی حجم زیادی از دیدگاه‌های متنوع و گاه متناقض را برای تصمیم‌گیری آگاهانه درباره خرید دشوار می‌کند. در این پژوهش، یک سامانه چندمرحله‌ای برای تحلیل خودکار و هوشمندانه دیدگاه‌های مشتریان با هدف سازمان‌دهی بخش نظرات ارائه شده است. این سامانه از مدل‌های زبانی بزرگِ همه‌منظوره بهره می‌گیرد. در مرحله نخست، سامانه پیشنهادی با برچسب‌گذاری نظرات، دیدگاه‌های مربوط به خدمات فروشگاه آنلاین را به‌صورت خودکار از نظراتی که مشخصاً درباره محصولات هستند، تفکیک می‌کند. سپس، جنبه‌های گوناگون محصولات را که در مجموع نظرات مطرح شده‌اند، استخراج می‌کند. در نهایت، نظرات را بر اساس جنبه‌های استخراج‌شده دسته‌بندی می‌کند. علاوه بر این، یک مجموعه‌داده جدیدِ برچسب‌گذاری‌شده و فارسی نیز به‌عنوان مجموعه‌داده معیار ایجاد شده است که شامل نظرات برچسب‌خورده مرتبط با فروشگاه‌های آنلاین است. نتایج آزمایش‌ها نشان می‌دهند که بهترین عملکرد متعلق به مدل Qwen 2.5 بوده است؛ این مدل به دقت ۹۱/۷ درصد دست یافته است.

کلمات کلیدی: دسته‌بندی نظرات، استخراج جنبه‌ها، مدل‌های زبانی بزرگ، تحلیل احساسات، سیستم‌های توصیه‌گر.