From Reviews to Results: Generative AI for Review-Driven Product and Service Comparisons

Cristian Cosentino[1 [0000−0002−6368−373X]],
Merve Gündüz Cüre[2 [0000−0002−7792−8056]],
Fabrizio Marozzo[1 [0000−0001−7887−1314]]✉, and
Şule Öztürk Birim[2 [0000−0001−7544−8588]]

[1] University of Calabria, Rende, Italy
{ccosentino, fmarozzo}@dimes.unical.it
[2] Manisa Celal Bayar University, Manisa, Turkey
{merve.gunduz, sule.ozturk}@cbu.edu.tr

Abstract. In today's digital world, user-generated reviews represent invaluable insights reflecting authentic experiences, preferences, and perceptions regarding products and services. Such reviews play a critical role both for consumers seeking informed purchasing decisions and businesses aiming to optimize their offerings and strategies. Recent advancements in machine learning, particularly the emergence of Large Language Models (LLMs) have significantly enhanced the processing and interpretation of this rich, unstructured textual data. Traditional platforms for comparing products and services focus primarily on structured specifications, often neglecting the detailed, experience-based information contained in user reviews. To address this limitation, we propose a novel framework that leverages Generative AI and advanced LLMs to extract and interpret user feedback, enabling more informed, experience-aware comparisons. Our approach involves three main phases: targeted review and metadata collection; topic modeling and sentiment classification using fine-tuned BERT models; and structured comparisons powered by user reviews, featuring attribute-level scores and natural language explanations generated by advanced GenAI tools such as GPT-4. Evaluated on real-world scenarios, including comparisons of similar Amazon products and nearby hotels, our framework outperforms traditional aggregation methods by generating more precise comparative scores and context-aware explanations.

Keywords: Large Language Models · Natural Language Processing · BERT · GPT · ChatGPT · Interpretable Models · Explainability

1 Introduction

In today's digital ecosystem, user-generated reviews have become indispensable, significantly influencing both consumer behavior and business strategy. Platforms such as Amazon, Booking.com, TripAdvisor, and Trustpilot host millions

of user opinions that capture real-world experiences, satisfaction levels, and individual preferences. This vast amount of textual feedback has been shown to shape purchasing decisions, build or damage brand trust, and affect market competitiveness. Beyond guiding consumers [5], reviews also provide organizations with actionable insights to improve product design, customer support, and marketing strategies [14]. Furthermore, recent studies suggest that analyzing sentiment and thematic patterns within user reviews can help predict product performance and identify emerging customer concerns before they escalate [15].

Despite their richness, user reviews remain underexploited in most product and service comparison systems. Traditional comparison platforms—such as shopping aggregators, travel booking sites, and consumer review portals—typically prioritize structured attributes like technical specifications, pricing, or expert scores, while ignoring the nuanced content of user-written reviews. When user input is included, it is often reduced to simplified summaries such as star ratings or sentiment scores, which fail to reflect the diversity and specificity of user experiences. Some leading platforms, including Amazon and Tripadvisor, have introduced concise summaries of user feedback for individual products or hotels, but they do not provide mechanisms to directly compare competing options based on this textual evidence.

To address these limitations, we introduce a unified framework that places user feedback at the core of product and service comparisons, enabling more meaningful and transparent evaluations. The process begins by systematically collecting reviews and accompanying structured data—such as technical specifications, product details, and helpfulness votes—for the specific products or services to be compared, sourced from the most relevant online platforms. We then apply BERTopic to extract dominant topics across the combined review corpus, and perform topic-specific sentiment analysis using fine-tuned BERT models to assess whether each theme is discussed positively, negatively, or neutrally for each product [2]. Finally, GPT-4 uses the topic–sentiment signals to generate structured comparison tables along with concise explanatory narratives that reflect both technical information and user opinions.

We conducted extensive experimental evaluations to assess the effectiveness of our review-informed comparison framework. The experiments involved multiple pairs of comparable Amazon products and Tripadvisor hotels, analyzed using two prompting strategies: a baseline that leverages technical information about the products or services, and an advanced method that integrates structured topic–sentiment information extracted from user reviews. Results show that the advanced approach produces more critical and user-aligned evaluations, particularly in identifying product weaknesses and supporting decisions. Evaluations confirm that the explanations generated through our framework are more nuanced and actionable. An ablation study further validates the contribution of each component—topic annotation, sentiment filtering, and structured prompting—by demonstrating consistent performance improvements as each element is introduced.

The remainder of the paper is structured as follows. Section 2 offers a concise review of related works. Section 3 describes the proposed framework. Section 4 presents in-depth comparisons of Amazon products and Tripadvisor hotels. Section 5 analyzes the impact of different prompting strategies on scores and explanations. Finally, Section 6 concludes the paper.

2 Related Work

Comparing products based on online reviews is essential for understanding customer preferences and guiding product development. Reviews offer rich, real-world insights into user satisfaction, expectations, and recurring issues. Prior studies have highlighted their value: Mudambi and Schuff [14] emphasized the role of big data in capturing user experience; Yang et al. [21] showed reviews can support satisfaction estimation; and Kessler et al. [11] proposed extracting comparative statements for product ranking.

Advancements in machine learning—particularly deep learning—have enabled automatic extraction of comparative insights. Arora et al. [1] a deep learning approach using LSTMs to extract product comparison information—product names, user opinions, and comparison aspects—from e-commerce reviews, while Sharma et al. [18] presents an RPA system that collects and processes e-commerce data to support personalized product recommendations and Li et al. [12] explored how e-WOM and review helpfulness influence user decisions.

Several frameworks have emerged to support structured product comparison. MatrixMiner [16] extracts feature-value pairs from unstructured text, and SRSS [10] combines topic detection and sentiment analysis to compare opinions. Recent work includes i-SPC [17], a Shopee-integrated tool using Naïve Bayes for multi-criteria evaluation; LDA-based sentiment-topic models for networked comparisons [9]; and web scraping systems for real-time product ranking [7]. Vedula et al.'s ReBARC [19] and HCPC [20] integrate catalog data with reviews to generate explainable, human-centered comparisons.

In addition to academic research, commercial platforms like *Consumer Reports*[3] and *CNET*[4] offer automated or semi-automated product comparisons. While widely used and informative—leveraging expert testing, editorial reviews, and user opinions—these services are typically limited to selected product categories and primarily rely on structured data and editorial insights, often lacking the granularity provided by topic-level sentiment analysis.

Our framework fills this gap by combining advanced topic modeling, sentiment classification, and natural language generation techniques to extract and synthesize insights from user reviews. By integrating with major online platforms and their APIs, it enables the comparison of virtually any product or service available online. The system filters out unreliable or outlier reviews to reduce sentiment distortion and constructs structured overviews that highlight strengths, weaknesses, and possible design rationales—directly grounded in user

[3] `https://consumerreports.org`
[4] `https://cnet.com`

experiences. Unlike traditional ranking systems, our approach supports in-depth, user-driven comparisons based on authentic feedback rather than curated specifications.

3 Proposed Framework

Our framework transforms technical specifications, consumer reviews, and platform metadata into structured insights that support informed product and service comparisons. It integrates classification and generation techniques within a three-phase pipeline, summarized in Figure 1.

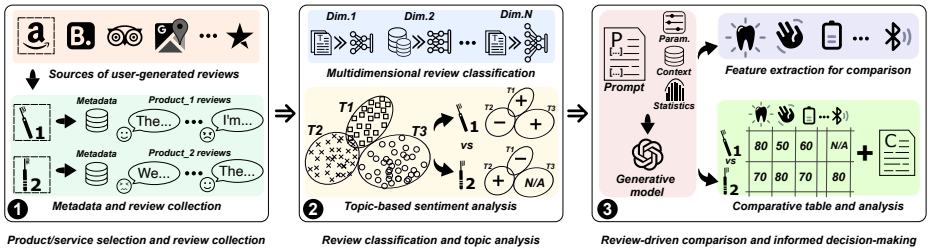


Fig. 1. Execution flow of the proposed framework.

The initial phase, *Product/service selection and review collection*, involves identifying relevant products or services and gathering consumer reviews from prominent platforms offering user-generated content, such as Amazon, TripAdvisor, Booking.com, and Trustpilot. Reviews are collected along with pertinent metadata—including product specification sheets and user feedback indicators (e.g., helpfulness ratings). To complement user opinions, technical specifications are also retrieved either directly from the web or inferred using language models when structured data is unavailable.

The second phase, *Review classification and topic analysis*, employs advanced language models, specifically fine-tuned BERT variants, to classify and analyze collected reviews. Initially, topic modeling techniques (e.g., BERTopic) identify dominant topics across the aggregated reviews. Subsequently, each review undergoes topic-specific sentiment analysis to ascertain whether user opinions related to a particular topic lean positively or negatively. This step includes rigorous outlier detection to ensure reliability; topics lacking sufficient review volume for specific products are excluded from detailed sentiment analysis, preventing unreliable conclusions based on sparse data.

In the final phase, *Review-driven comparison and informed decision-making*, the framework first defines the comparative features based on the most salient topics identified across the review corpus. It then combines technical information and model knowledge with aggregated statistics and individual review annotations—including topic and sentiment labels—to assign scores and generate

structured comparison tables. These tables highlight how each product or service is evaluated on key aspects by real users, supporting transparent and evidence-based decision-making. The framework also interprets user criticisms in the context of potential design trade-offs, offering explanations for recurring issues. This approach supports a more grounded understanding of strengths and limitations as perceived by users.

3.1 Implementation Details of Classification Algorithms and Topic detection

Each review in our pipeline is annotated along two orthogonal dimensions: *sentiment* and *topic*. To compute sentiment scores, we use the pretrained transformer-based model `siebert/sentiment-roberta-large-english`[5], developed by Hartmann et al.[8]. This model is based on the RoBERTa-large architecture[13] and was fine-tuned on a diverse corpus of sentiment-labeled texts, including tweets and customer reviews. Hartmann et al. [8] show that the model outperforms common sentiment lexicons by 20% and achieves over 15% higher accuracy than a DistilBERT model trained on SST-2.

For topic extraction, we adopted *BERTopic*[6], which outperformed alternative methods in both consistency and diversity of topics. The model clusters review embeddings—generated using a pretrained RoBERTa model—into coherent thematic groups. The number of topics is selected by maximizing topic coherence, following the protocol of Chen et al.[3], to balance granularity and generality. Each review is then assigned a dominant topic label, corresponding to the theme with the highest prominence score.

Once topics are identified, we aggregate sentiment within each topic to construct a structured representation of user opinion. Specifically, for each product/service, we compute a topic–sentiment vector containing one entry per shared topic, where each value represents the aggregated sentiment of all reviews associated with the corresponding topic, expressed on a five-level scale (*strongly negative*, *slightly negative*, *neutral*, *slightly positive*, *strongly positive*). This representation enables the comparison of multiple products (two or more) within a unified topic space. By preserving sentiment granularity, the framework captures subtle differences in user perception across competing items. These structured vectors are then used to condition the prompt, guiding the generation of concise and balanced comparative narratives informed by topic-level sentiment signals and representative review excerpts.

3.2 Implementation Details of Prompt Strategies

We evaluate four prompts of increasing complexity, each adding more contextual information to guide the model's judgments. While our implementation relies on OpenAI-4o accessed via the API, the same framework is compatible with other generative models such as Gemini or Claude, provided they support structured

[5] `https://huggingface.co/siebert/sentiment-roberta-large-english`

Prompt	Product names	Generate comp. features	Reviews	Topics	Sentiment +outliers	Explanation
`base`	✓	✓				
`intermediate1`	✓	✓	✓			
`intermediate2`	✓	✓	✓	✓		
`advanced`	✓	✓	✓	✓	✓	✓

Table 1. Prompt variants and included input components. Checkmarks indicate which elements are provided at each stage of prompt refinement.

input and instruction-following capabilities. Table 1 summarizes the input components provided with each prompt.

The simplest variant, referred to as the `base` prompt, relies solely on the model's general knowledge and does not incorporate any additional user-provided information. It provides the names of multiple products or services stored in the variable `$products`, and instructs the model to generate a list of dictionaries—each evaluating a specific *product-feature* pair with a *score* in the range `[1, 5]` and a *description*. While the prompt supports comparisons involving any number of products, our experiments focus on scenarios involving two products or services. Although prompts can generate features directly, this often leads to inconsistent or non-overlapping sets across items, hindering comparability. To ensure stable and aligned evaluations, we extract features externally—combining domain knowledge, topic modeling, and product specifications—and supply them as structured input to the prompt.

> `Base prompt`: Your task is to evaluate and compare multiple products listed in {`$products`} using a predefined set of features given in {`$features`}.
>
> For each product and each feature, generate an individual evaluation in the form of a dictionary with the following fields: `product` (a short name or identifier of the product), `feature` (the name of the evaluated feature), `score` (a numerical rating in the range `[1, 5]`), and `description` (a concise and objective justification for the assigned score). The response must be a list of such dictionaries, each evaluating a single product-feature pair.
>
> Do not include any introductory or summary text. Do not compare products directly within descriptions. Ensure that each dictionary is self-contained and independently describes the evaluation.

To improve grounding, the next variant—`intermediate1`—supplements the prompt with the full set of user reviews for each product. This allows the model to rely on authentic customer feedback rather than general knowledge alone. Building on this, `intermediate2` introduces the dominant topic labels discovered by BERTopic from the reviews, helping to focus the model's attention on the most salient and recurrent themes.

A further enhancement in the `advanced` prompt is the structured representation of each review as a JSON object, including the review text, its sentiment, and its dominant topic. This enriched format offers a more granular view of product perception across topics. The prompt also incorporates aver-

age *topic–sentiment* distributions to guide the model in generating feature-level evaluations. The model considers both the presence and polarity of each topic and receives explicit instructions for handling missing or sparse information.

> `Advanced prompt`: Your task is to evaluate and compare multiple products listed in {`$products`} using a predefined set of features in {`$features`}. For each product, {`$<product>_reviews`} supplies a list of enriched reviews, each represented as a JSON object with the fields {*text*, *sentiment*, *topic*}. Aggregate data are also provided for every `<product>`, `<topic>` pair:
>
> – {`$<product>_<topic>_sentiment`} — overall sentiment label (e.g., `high negative`, `low positive`);
> – {`$<product>_<topic>_info`} — number of reviews in that topic.
>
> For every feature of every product, combine (i) technical knowledge, (ii) the review texts, and (iii) the aggregate statistics. Produce a list of dictionaries; each dictionary contains:
>
> – `Product` – product name;
> – `Feature` – feature name;
> – `Score` – a value in `[1,5]` or `"N/A"` if the feature is absent;
> – `Description` – a concise explanation blending technical and review-based insights.
>
> Do not include introductory or summary text. Do not compare products within descriptions. Ensure each dictionary is self-contained and independently explains its evaluation.

Finally, we introduce an auxiliary prompt that asks the model to explain low-scoring or negatively weighted features by identifying plausible causes—such as design trade-offs, usability flaws, or service gaps—thereby enhancing the interpretability of the generated evaluations.

4 Experimental Results

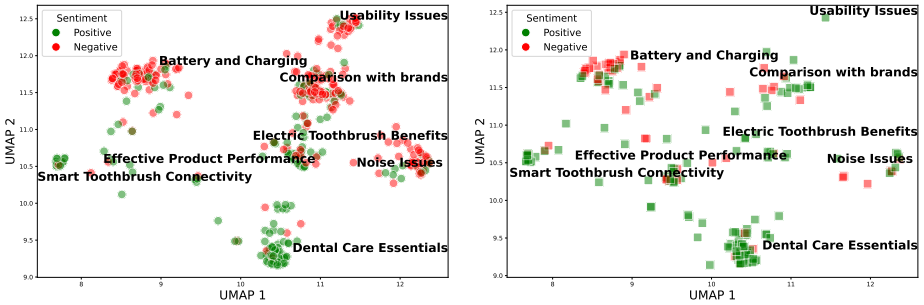
The experimental evaluation of our framework is based on two curated datasets, publicly available at `https://github.com/SCAlabUnical/UserReviewDatasets`, designed to support comparative analysis of similar products and services using user-generated reviews. The first dataset contains approximately 10,000 Amazon reviews of functionally equivalent electronic items (e.g., electric toothbrushes, USB cables), with structured metadata such as ratings, titles, verification status, and timestamps. The second dataset includes Tripadvisor reviews of hotels in New York City, with rich metadata such as sub-ratings (e.g., service, value), trip types, and management responses. It comprises 150 hotels and several hundred reviews per hotel. Together, these datasets enable systematic, feature-level comparisons grounded in both user sentiment and thematic relevance.

4.1 Use Case: In-Depth Comparison of Two Electric Toothbrushes via Amazon Reviews

For demonstration purposes, we illustrate our framework using two electric toothbrushes available on Amazon, referred to here as `Model-A` and `Model-B`, as a case study. Although these represent real commercial products, we anonymize their identities to maintain neutrality, avoid brand bias, and focus solely on the methodological aspects of the review-based analysis.

`Model-A` is the newer-generation, professional-grade toothbrush, designed with a more compact and ergonomic form. It features a hard plastic body that gives it a modern, durable look and feel. In addition to improved cleaning performance, it includes a built-in pressure sensor and timer, catering to users who value efficiency and simplicity in oral hygiene. `Model-B`, by contrast, belongs to the previous product generation. It features a more classic design, is slightly bulkier in size, and uses smooth plastic for a softer tactile finish. Its main differentiator is the inclusion of Bluetooth connectivity—specifically in the version of `Model-B` analyzed—which enables integration with a mobile app for brushing feedback and usage tracking.

These two products, which differ in design era, materials, form factor, and functional focus, provide an ideal setting to demonstrate our framework's ability to extract and compare multidimensional insights from Amazon reviews. Specifically, we analyze sentiment and referenced topics to reveal how users perceive and evaluate each product.



(a) Model-A. (b) Model-B.

Fig. 2. UMAP visualization of BERTopic clusters for `Model-A` and `Model-B` reviews, with sentiment polarity (green = *positive*, red = *negative*) and topic labels generated from representative keywords.

We performed sentiment classification for each review and extracted discussion topics using BERTopic. Figures 2(a) and 2(b) show the topic clusters for `Model-A` and `Model-B`, respectively, visualized in a two-dimensional latent space using UMAP. Each cluster represents a distinct discussion theme, with topic names generated by ChatGPT based on the top representative keywords in each group. Green points denote *positive reviews*, while red points indicate *negative ones*. Finally, we assign each cluster an average sentiment label on a five-level scale (*strongly negative*, *slightly negative*, *neutral*, *slightly positive*, or *strongly positive*) to consistently capture differences in user sentiment across topics.

For `Model-A`, the sentiment distribution across topics is relatively balanced. Clusters such as *Effective Product Performance* and *Dental Care Essentials* exhibit a predominance of positive reviews. In contrast, negative sentiment is more

concentrated in topics like *Battery and Charging*, *Usability Issues*, *Comparison with Brands*, and *Noise Issues*. `Model-B`, on the other hand, shows generally less negative sentiment across most clusters. However, negative feedback still appears, particularly within *Usability Issues*, *Comparison with Brands*, and *Noise Issues*, albeit with slightly lower intensity than in `Model-A`.

	Base		Advanced	
Feature	Model-A	Model-B	Model-A	Model-B
Battery and Charging	4.0	4.5	3.0	3.5
Comparison with brands	4.0	4.5	3.5	4.5
Dental Care Essentials	4.0	4.5	4.0	4.0
Noise Issues	3.5	4.0	3.5	4.0
Electric Toothbrush Benefits	4.5	4.5	4.5	4.5
Usability Issues	4.0	4.5	2.5	4.0
Effective Product Performance	4.5	4.5	4.5	4.0
Smart Toothbrush Connectivity	1.0	5.0	N/A	3.5

Table 2. Evaluation scores (1–5 scale) for `Model-A` and `Model-B` using the `base` and `advanced` approaches (1 = very poor, 5 = excellent).

Table 2 shows the average evaluation scores for `Model-A` and `Model-B`, computed using the `base` and `advanced` prompting approaches. Scores are on a 1–5 scale, following a Likert-style evaluation where 1 indicates very poor performance and 5 denotes excellent performance. Under the `base` approach, `Model-B` exceeds `Model-A` on nearly every feature, driven chiefly by its Bluetooth connectivity and higher usability rating. In the `advanced` evaluation—where review evidence tempers the purely technical view—the gap narrows: `Model-B` still leads overall, but several of its ratings drop once user criticisms are considered, while `Model-A` retains a clear edge in *Effective Product Performance*. The `advanced` scores therefore surface trade-offs (e.g., connectivity reliability, grip comfort) that the `base` summary overlooks, giving a more critical, usage-oriented picture of both brushes.

Due to space constraints, we do not include full feature descriptions in the comparison table. Instead, we focus on those aspects where evaluation differences between the two models are most significant. For example, in the case of *Usability Issues*, notable contrasts emerge in both the `base` and `advanced` evaluations:

`[Model-A] Usability Issues description (base)` = Simple one-button interface is very intuitive but provides no feedback on pressure level or brushing time beyond the built-in timer. [Score = 4.0]
`[Model-B] Usability Issues description (base)` = Offers LED mode/pressure feedback, but initial Bluetooth setup can be cumbersome and may confuse non-tech-savvy users. [Score = 4.5]

`[Model-A] Usability Issues description (advanced)` = Intuitive design, but lacks interactive feedback. Frequent user complaints include slippery handle, stiff brush heads, awkward button layout, and unclear pressure signals, making daily use frustrating for

many. *These issues likely stem from design trade-offs: the hard-plastic shell may improve waterproofing and hygiene by avoiding textured surfaces or seams, but this reduces grip and tactile comfort.* [Score = 2.5]
[Model-B] Usability Issues description (advanced) = Offers LED guidance but has complex setup. Users do not often mention setup problems, but the Bluetooth connection and settings may still be confusing for people not used to technology. [Score = 4.0]

A second example involves *Smart Toothbrush Connectivity*:

[Model-A] Smart Toothbrush connectivity description (base) = No Bluetooth or app connectivity available, limiting interactive guidance and progress tracking. [Score = 1.0]
[Model-B] Smart Toothbrush connectivity description (base) = Bluetooth connectivity with the official app provides real-time brushing tracking, coaching, and personalized reports. [Score = 5.0]

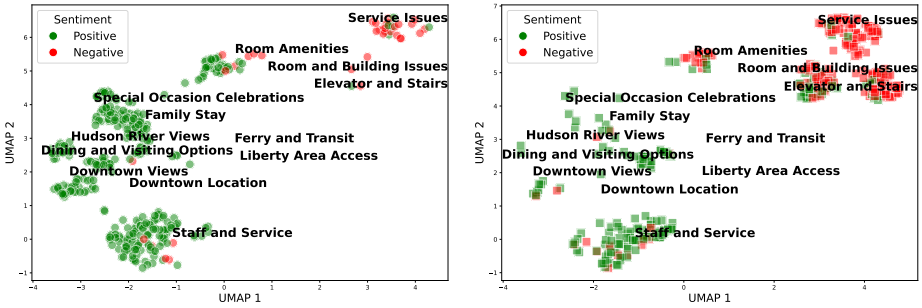
[Model-A] Smart Toothbrush connectivity description (advanced) = Lacks Bluetooth or connectivity features. Not discussed in reviews due to absence. [Score = N/A]
[Model-B] Smart Toothbrush connectivity description (advanced) = Offers app-based brushing feedback and reports. Reviews praise guidance but mention frequent disconnects and feedback gaps. [Score = 3.5]

The paired excerpts for *Usability Issues* and *Smart Toothbrush Connectivity* show why the review-aware advanced prompt is more diagnostic than the technical-only base version. In the first case, the `base` summaries simply note an intuitive one-button layout for *Model A* and LED feedback with a fussy Bluetooth setup for *Model B*. The `advanced` summaries, however, bring the user voice to the forefront by detailing the slippery handle, stiff brush heads, awkward button placement and unclear pressure signals that frustrate *Model A* owners, while also acknowledging that *Model B*'s guidance works yet can confuse the less tech-savvy. In the second case, the `base` layer records the presence or absence of Bluetooth and assigns extreme scores, but the advanced layer adds context: *Model B*'s app is applauded for coaching yet criticised for frequent disconnects, and *Model A* is marked "N/A" rather than a very low value to signal that the feature is missing, not implemented. By incorporating user experiences and highlighting possible design trade-offs, the `advanced` approach helps contextualize complaints and offers a more detailed and balanced foundation for comparison.

4.2 Use Case: Comparative Analysis of Two Downtown Hotels via TripAdvisor Reviews

To demonstrate the portability of our framework to the hospitality domain, we analyse two large New-York-City properties—hereafter `Hotel-A` (a riverside all-suite brand) and `Hotel-B` (a business-class high-rise belonging to the same global chain). Both are located within Lower Manhattan's financial district and share similar price points, yet they differ in age, room configuration and service philosophy, making them a suitable pair for review-based comparison.

Figures 3(a) and 3(b) show the UMAP projection of BERTopic clusters extracted from the TripAdvisor reviews. For `Hotel-A` (Fig.3(a)), the majority of clusters are dominated by positive sentiment. However, negative feedback is



(a) Hotel-A. (b) Hotel-B.

Fig. 3. UMAP visualization of BERTopic clusters for `Hotel-A` and `Hotel-B` reviews, with sentiment polarity (green = positive, red = negative) and topic labels generated from representative keywords.

concentrated in a compact *Service Issues* cluster in the upper-right corner. In contrast, `Hotel-B` (Fig.3(b)) exhibits a broader and more dispersed pattern of negative sentiment. Several clusters—most notably *Elevator and Stairs*, *Service Issues*, *Room Amenities*, and *Room and Building Issues*—show a higher concentration of negative reviews, indicating recurring dissatisfaction across multiple dimensions of the guest experience. This distribution suggests more widespread critical feedback, despite the otherwise similar thematic structure.

	Base		Advanced	
Feature	Hotel-A	Hotel-B	Hotel-A	Hotel-B
Staff and Service	4.5	4.0	4.5	4.0
Service Issues	5.0	3.0	3.0	2.0
Family Stay	3.5	4.0	4.0	4.0
Elevator and Stairs	4.5	4.0	4.5	2.5
Room Amenities	5.0	4.0	5.0	3.0
Room and Building Issues	4.5	4.0	4.5	3.0
Downtown Location	5.0	5.0	5.0	5.0
Hudson River Views	4.5	4.0	4.5	4.0
Downtown Views	4.5	4.0	4.5	4.0
Special Occasion Celebrations	4.5	4.0	4.5	4.0
Liberty Area Access	5.0	5.0	5.0	5.0
Ferry and Transit	5.0	4.5	5.0	4.5
Dining and Visiting Options	4.5	4.5	4.5	4.5

Table 3. Evaluation scores (1–5 scale) for `Hotel-A` and `Hotel-B` using the `base` and `advanced` approaches (1 = very poor, 5 = excellent).

Table 3 reports 1–5 scores for each feature under two evaluation schemes: a `base` layer that covers tangible facilities and location descriptors, and an `advanced` layer enriched with sentiment-weighted insights from guest narratives.

Three features illustrate how the review layer reshapes the assessment. For *Elevator and Stairs*, both hotels earn strong `base` marks for having modern lifts, yet `Hotel-B`'s `advanced` score falls sharply once guests' complaints about long waits and crowding are considered, whereas `Hotel-A` remains unchanged. In *Service Issues*, operational metrics give `Hotel-A` a perfect score, but the review layer lowers that value and pushes `Hotel-B` even lower after noting delays, limited breakfast hours, and patchy room service. Conversely, the *Family Stay* score for `Hotel-A` rises after positive family experiences in the reviews offset its modest `base` value, while `Hotel-B` remains stable. These shifts show how the descriptive layer grounds the numbers in real usage contexts and reveals friction points that raw specifications alone can overlook.

5 Performance Comparison

To assess explanation quality across domains, we applied both prompting strategies—`base` and `advanced`—to ten Amazon product pairs and ten hotel pairs. For each pair, we generated a feature-comparison table for both items under both strategies. These tables contain attribute-level scores as well as concise descriptions that justify each score, and the evaluation focuses on these explanations rather than on the products themselves. This automated "LLM-as-a-judge" protocol is now common in the literature because it scales well, enforces a consistent rubric, and correlates closely with expert judgment when manual evaluation is impractical [22,4].

Two reasoning-oriented language models, *OpenAI-o3* and *DeepSeek-R1*, served as independent evaluators. Each model rated every explanation on a five-point Likert scale (1 = very poor, 5 = excellent) according to five criteria relevant to product and service comparison:

- *Feature Coverage*: extent to which all salient characteristics are mentioned;
- *Strength Identification*: clarity and completeness in highlighting advantages;
- *Weakness Identification*: accuracy and balance in exposing flaws or trade-offs;
- *Comparative Insight*: effectiveness in contrasting the item with its alternatives;
- *Decision Support*: usefulness of the explanation in guiding a justified choice.

Table 4 confirms the advantage of the review-aware `advanced` prompt. Across both datasets and for both evaluators, the advanced explanations obtain higher scores on the three dimensions most relevant to decision-making—*Weakness Identification*, *Comparative Insight*, and *Decision Support*. This indicates that injecting review evidence makes the text more critical and more useful when weighing alternatives. Results for *Feature Coverage* and *Strength Identification* are mixed. With *OpenAI-o3*, the advanced prompt covers features slightly better, whereas *DeepSeek-R1* shows no clear gain and even a small edge for the base version in a few cases. Strength identification also favours the base summaries in most Amazon pairs and under *DeepSeek-R1* overall. These modest reversals

Evaluator	Metric	Amazon dataset		Tripadvisor dataset	
		Base	**Adv**	**Base**	**Adv**
OpenAI-o3	Feature Coverage	4.1	**4.4**	4.1	**4.5**
	Strength Identification	**4.2**	4.1	4.0	**4.1**
	Weakness Identification	2.6	**4.5**	2.5	**4.3**
	Comparative Insight	3.5	**4.1**	3.6	**4.2**
	Decision Support	3.8	**4.2**	3.9	**4.3**
	Mean	3.5	**4.3**	3.6	**4.3**
DeepSeek-R1	Feature Coverage	**4.1**	4.1	**4.2**	4.1
	Strength Identification	**4.2**	4.0	**4.1**	4.0
	Weakness Identification	3.2	**5.0**	3.6	**4.9**
	Comparative Insight	4.0	**4.5**	4.0	**4.4**
	Decision Support	3.9	**5.0**	3.9	**4.8**
	Mean	3.9	**4.5**	4.0	**4.4**

Table 4. Average LLM-as-a-judge scores (1–5) for Amazon product pairs and Tripadvisor hotel pairs, judged by *OpenAI-o3* and *DeepSeek-R1*. Bold marks each evaluator's best metric.

are expected: when the prompt devotes more space to user criticism, it can leave fewer tokens for exhaustive feature lists or positive framing. Even so, the advanced prompt consistently surfaces user-reported issues more clearly and offers stronger comparative guidance, outweighing the small loss in upbeat content.

The values reported in Table 4 are means computed from ten fully independent runs for each evaluator, with each run executed in isolation to avoid cross-session contamination. The results were stable across repetitions, with low run-to-run variance observed for both base and advanced prompts. Paired-samples t-tests conducted on the per-run scores revealed that advanced prompts consistently outperformed base prompts across most dimensions ($p < 0.05$), confirming that the observed differences were statistically significant across runs.

Table 5 reports composite mean scores for each prompt variant. As additional review-grounded signals are introduced—moving from the `base` prompt, through two intermediate variants, to the fully review-aware `advanced` prompt—the quality of the generated explanations generally rises. *OpenAI-o3* improves monotonically on both datasets, while *DeepSeek-R1* shows a brief dip at `inter1` on the Amazon items before climbing again. The overall pattern is clear: richer inputs that blend topic labels and sentiment statistics help the models produce more

Dataset	Evaluator	Base	Inter1	Inter2	Advanced
Amazon	OpenAI-o3	3.5	3.8	3.9	4.3
	DeepSeek-R1	3.9	3.9	4.2	4.5
Tripadvisor	OpenAI-o3	3.6	3.8	4.0	4.3
	DeepSeek-R1	4.0	4.1	4.2	4.4

Table 5. Composite ablation scores. Values are the mean of the five LLM-as-a-judge criteria for each prompt variant, computed per dataset and evaluator.

informative and actionable comparisons, with the `advanced` prompt delivering the highest scores across evaluators and domains.

6 Conclusions

In today's digital landscape, user-generated reviews are a rich source of consumer insight that increasingly shapes purchasing behavior, product design, and business strategy. While modern language models such as BERT and GPT have enabled more powerful tools for extracting information from this unstructured content, effective comparison still requires going beyond raw classification or rating aggregation.

Our framework addresses this need by combining topic-specific sentiment analysis with structured generative summarization to deliver more accurate, balanced, and informative comparisons. By grounding evaluations in both user feedback and product metadata, the system supports decision-making with context-aware insights rather than opaque scores. Results across both Amazon and Tripadvisor domains show that review-aware prompting consistently outperforms traditional specification-based methods. The ablation analysis further highlights the incremental benefits of topic annotation, sentiment filtering, and structured input design.

The architecture is modular and can evolve with new inputs or deployment contexts. Future directions include incorporating multimodal signals (e.g., images or video), adapting to real-time review streams, and supporting interactive or user-personalized comparison tools. Addressing challenges such as opinion bias, review sparsity, and domain transfer will also be essential to ensure the robustness and scalability of the approach.

Acknowledgements and Competing Interests

This work was supported by the research project "INSIDER: INtelligent ServIce Deployment for advanced cloud-Edge integRation" granted by the Italian Ministry of University and Research (MUR) within the PRIN 2022 program and European Union - Next Generation EU (grant n. 2022WWSCRR, CUP H53D23003670006). The authors declare that they have no competing interests.

References

1. Arora, J., et al.: Extracting entities of interest from comparative product reviews. In: CIKM. pp. 1975–1978 (2017)
2. Cantini, R., Cosentino, C., Marozzo, F.: Multi-dimensional classification on social media data for detailed reporting with large language models. In: 20th Int. Conference on Artificial Intelligence Applications and Innovations. pp. 100–114 (2024)
3. Chen, Y., Peng, Z., Kim, S.H., Choi, C.W.: What we can do and cannot do with topic modeling: A systematic review. Communication Methods and Measures **17**(2), 111–130 (2023)

4. Chiang, C.H., Lee, H.y.: Can large language models be an alternative to human evaluations? arXiv preprint arXiv:2305.01937 (2023)
5. Cosentino, C., Gündüz-Cüre, M., Marozzo, F., Öztürk-Birim, Ş.: Exploiting large language models for enhanced review classification explanations through interpretable and multidimensional analysis. In: International Conference on Discovery Science. pp. 3–18. Springer (2024)
6. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure (2022)
7. Harikirshnan, K., et al.: Intelligent online shopping using ml-based product comparison engine. 6th International Conference on Inventive Computation Technologies, ICICT 2023 - Proceedings pp. 174–179 (2023)
8. Hartmann, J., Heitmann, M., Siebert, C., Schamp, C.: More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing **40**(1), 75–87 (2023)
9. He, Z., Zheng, L., He, S.: A novel approach for product competitive analysis based on online reviews. Electronic Commerce Research **23**, 2259–2290 (12 2023)
10. Jin, J., Ji, P., Yan, S.: Comparison of series products from customer online concerns for competitive intelligence. Journal of Ambient Intelligence and Humanized Computing **10**, 937–952 (3 2019)
11. Kessler, W., Klinger, R., Kuhn, J.: Towards opinion mining from reviews for the prediction of product rankings. In: WASSA. pp. 51–57 (2015)
12. Li, Y., Zheng, J., Yue, S., Zhi-ping, F.: Capturing and analyzing e-wom for travel products: a method based on sentiment analysis and stochastic dominance. Kybernetes **51**, 3041–3072 (2021)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 (2019)
14. Mudambi, S.M., Schuff, D.: Research note: What makes a helpful online review? a study of customer reviews on amazon. com. MIS quarterly pp. 185–200 (2010)
15. Myers, D., et al.: Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. Cluster Computing **27**(1), 1–26 (2024)
16. Nasr, S.B., et al.: Automated extraction of product comparison matrices from informal product descriptions. Journal of Systems and Software **124**, 82–103 (2017)
17. Samah, K.A.F.A., et al.: Intelligence shopee product comparison (i-spc) and visualization of product information via naïve bayes adaptation. Journal of Advanced Research in Applied Sciences and Engineering Technology **37**, 179–190 (7 2024)
18. Sharma, D.K., Lohana, S., Arora, S., Dixit, A., Tiwari, M., Tiwari, T.: E-commerce product comparison portal for classification of customer data based on data mining. Materials Today: Proceedings **51**, 166–171 (2022)
19. Vedula, N., Collins, M., Agichtein, E., Rokhlenko, O.: What matters for shoppers: Investigating key attributes for online product comparison. AI and Lecture Notes in Bioinformatics) pp. 231–239 (2022)
20. Vedula, N., Collins, M., Agichtein, E., Rokhlenko, O.: Generating explainable product comparisons for online shopping. WSDM 2023 pp. 949–957 (2 2023)
21. Yang, C., Wu, L., Kun, T., Yu, C., Zhou, Y., Tao, Y., Song, Y.: Online user review analysis for product evaluation and improvement. Journal of Theoretical and Applied Electronic Commerce Research **16**, 1598–1611 (2021)
22. Zhang, X., Li, Y., Wang, J., Sun, B., Ma, W., Sun, P., Zhang, M.: Large language models as evaluators for recommendation explanations. In: 18th ACM Conf. on Recommender Systems. pp. 33–42 (2024)