Balanced and Token-Efficient Summarization of User Reviews via Stratified Sampling and Large Language Models

Fabrizio Marozzo[1 [0000−0001−7887−1314]] (✉), Loris Belcastro[1 [0000−0001−6324−8108]], Cristian Cosentino[1 [000−0002−6368−373X]], and Pietro Liò[2 [000−0002−6368−373X]]

[1] University of Calabria, Rende 87036, Italy
{fmarozzo,lbelcastro,ccosentino}@dimes.unical.it
[2] University of Cambridge, Cambridge CB3 0FD, United Kingdom
pietro.lio@cl.cam.ac.uk

Abstract. User-generated reviews offer valuable insights into consumer experiences, preferences, and concerns. They provide direct feedback on product perception and improvements while helping users evaluate strengths, weaknesses, and alternatives. Advanced machine learning techniques, including LLMs like BERT and GPT, enhance the extraction of meaningful information from these vast datasets. This paper introduces a framework leveraging Large Language Models (LLMs) to generate high-quality summaries using minimal input tokens. By employing multidimensional classification (sentiment, topics, emotion) combined with a stratified sampling approach, our framework selects a compact yet comprehensive subset of reviews that accurately represents the original dataset. Tailored prompts guide the LLMs to create balanced summaries that fairly represent both strengths and weaknesses. Experiments on Amazon and Tripadvisor datasets demonstrate that our method significantly reduces token usage and computational costs, while consistently outperforming traditional AI-based summarization approaches in terms of content coverage, balance, and semantic accuracy.

Keywords: Large Language Models · Generative AI · AI-Generated Summaries · Review Aggregation · Opinion Mining

1 Introduction

In the current digital landscape, user-generated reviews provide essential insights into consumer experiences, preferences, and concerns across various industries. Businesses leverage these reviews to evaluate product performance, identify areas needing improvement, and better adapt to consumer expectations [26]. Simultaneously, consumers rely on reviews to assess product strengths, weaknesses, and available alternatives before making purchasing decisions [30].

However, effectively extracting meaningful insights from extensive volumes of user-generated content requires advanced classification techniques, such as sentiment analysis and topic modeling [18]. Once classified, reviews can be grouped

and filtered to select the most representative examples. While platforms like Amazon and Booking display relevant reviews to aid consumers, ensuring fairness and representativeness in review selection is challenging but critical. Representative reviews must clearly highlight a product's key attributes, including both positive and negative aspects, as well as distinctive features [12]. Recent advances in Large Language Models (LLMs) offer significant potential to enhance this classification and selection process, improving the quality, relevance, and balance of insights extracted from reviews [23].

In this paper, we propose a novel framework leveraging LLMs to systematically classify and summarize user-generated reviews, ensuring balanced and comprehensive insights while significantly reducing computational costs. Our approach first classifies reviews across multiple dimensions (sentiment, topics, and emotions) and then employs a stratified sampling strategy to select a compact yet representative subset. This carefully constructed sample accurately mirrors the original dataset distribution across all dimensions. Subsequently, a generative LLM processes the selected subset, guided by tailored prompts, to produce balanced summaries that fairly represent both positive and negative viewpoints without bias. Crucially, our framework prioritizes token efficiency by using minimal input tokens, significantly enhancing scalability and cost-effectiveness for processing large review datasets [20].

To validate the proposed framework, we conducted comprehensive experiments on Amazon and Tripadvisor datasets, representing diverse consumer sentiments and product categories. Results demonstrate that our method consistently produces summaries of superior quality compared to conventional AI summarization approaches, as evidenced by text quality scores, latent semantic representations, and evaluations from automated tools and human assessors. Additionally, our compact and representative sampling strategy substantially reduces token usage and computational resources without compromising summary quality, thereby achieving optimal efficiency and scalability.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 presents the proposed framework. Section 4 discusses the results. Finally, Section 5 concludes the paper.

2 Related work

Artificial intelligence powered by LLMs has revolutionized data extraction, enabling faster, more intuitive processing of natural language queries for report generation, question answering, and data visualization. These models are widely applied in education, e-commerce, healthcare, and entertainment, particularly through chatbots for information retrieval [4]. In education, they assist in lesson planning [17], while in healthcare, they aid in disease detection and diagnosis [3]. In e-commerce, LLMs enhance customer support and shopping experiences [10] and outperform human crowd-workers in data annotation tasks [8].

LLMs like GPT are effective in generating descriptive summaries across domains, from extracting insights from tables [24] to summarizing medical re-

ports [15] and financial texts [28]. In cybersecurity, they help summarize system logs, improving data organization and security audits [31]. However, evaluating LLM-generated summaries remains challenging due to semantic quality concerns and hallucinations [22]. To address these issues, a systematic evaluation framework was proposed, assessing reports based on completeness, accuracy, verifiability, and responsiveness to information needs [14].

LLMs have also enhanced the analysis of content posted on social media through sentiment analysis, topic modeling, and summarization [6], supporting applications like product recommendation and market research. Sentiment classification has been improved using CNN-based functions [2], hybrid LSTM-CNN models for tweets [19], and GRU-based models for product reviews [1]. Comparative studies [21] show that models like GPT-3.5 and LLaMA-2 excel in predicting ratings and understanding sentiment. However, evaluating the quality of LLM-generated summaries remains a complex task, as standard metrics like TF-IDF, ROUGE-L, or S-BERT capture only limited aspects of informativeness and coherence. For example, [7] propose a threefold evaluation combining semantic metrics, LLM-as-evaluator scoring, and expert judgment. Similarly, [11] introduce fairness-aware measures like Equal Coverage and Coverage Parity to assess representation across social attributes while accounting for redundancy. Despite these advances, reliably capturing the semantic quality and fairness of LLM outputs remains an open challenge.

In contrast to previous work, our study categorizes user reviews across sentiment, emotion, and topic to improve the accuracy and completeness of LLM-generated reports. Moreover, it aims to ensure fair and balanced summaries, providing consumers with comprehensive insights that reflect both positive and negative feedback. By employing multidimensional classification and stratified sampling, we effectively capture consumer priorities. Additionally, we introduce a robust evaluation framework, benchmarking report quality through quantitative text metrics, latent representations, and qualitative assessments via automated tools and human evaluations.

3 Proposed framework

The proposed framework analyzes online reviews, classifies sentiments, and extracts key topics to generate balanced and detailed reports. These reports highlight strengths and weaknesses, aiding consumers and helping companies refine their offerings. As shown in Figure 1, the framework comprises four main phases, which are detailed below.

The initial phase consists of systematically collecting *product/service reviews* from *online platforms*, through the use of official APIs or API of sites specialized in downloading data. Platforms provide access to products, services, and user reviews, such as Amazon for products, Booking and Tripadvisor for hotels, Tripadvisor and Google Maps for restaurants, and TrustPilot for services. Once you have chosen a platform, chosen a product/service, user reviews are downloaded together with all the metadata of the reviews themselves (e.g. how many people

found a review useful) and of the product (e.g. information sheet). This process in some cases requires the use of specific keywords or filters on the reviews to ensure that the collected reviews are directly related to consumer experiences. To this end, the framework provides a filtering mechanism to select only the most relevant reviews, thus laying a solid foundation for the targeted and efficient analysis performed in the subsequent phases.

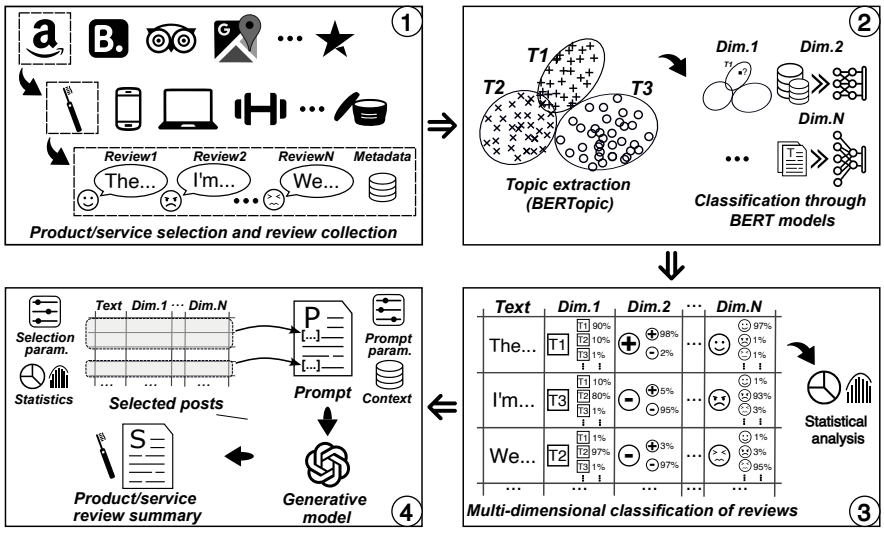


Fig. 1. Execution flow of the proposed framework.

The second phase utilizes analytical LLMs for multidimensional classification, employing tools like optimized BERT models to classify reviews across dimensions such as sentiment (positive or negative), emotion (e.g., joy, anger), and topics [5]. Each review is analyzed to determine its overall sentiment, the emotions expressed by users, and the topics discussed. For topic analysis (*BERTopic*), a comprehensive review of all topics must first be conducted, enabling the model to identify and explain which topic is addressed in a given review. For other dimensions, classification models (e.g., *BERT models*) are trained on specific datasets to accurately identify dimensions and their respective classes. These models not only assign labels (e.g., positive or negative sentiment) but also provide probabilities (e.g., 95% positive, 5% negative), ensuring nuanced and reliable classification.

The third phase organizes all collected reviews, along with their metadata and classifications from the different examined dimensions (*multi-dimensional classification*). This classification enables the grouping of similar reviews across dimensions such as sentiment (positive and negative), topic (various discussion topics), emotions (e.g., happy, sad), and time (e.g., last month, this year), providing a comprehensive understanding of consumer feedback. *Statistical data*

analysis is conducted to examine the distribution of reviews across all these dimensions, offering valuable insights into the prevalence and relationships of sentiments, topics, emotions, and temporal trends. This comprehensive process is fundamental to improving the completeness and depth of the analysis.

The final step of our framework utilizes generative LLMs to create human-readable summaries of product/service reviews that are both comprehensive and balanced, capturing their strengths as well as their weaknesses. The process leverages the results of multidimensional classification, where a stratified sampling method selects a compact, representative sample that reflects the distribution of all dimensions and classes in the original dataset (e.g., positive and negative aspects for the sentiment dimension) while minimizing token usage. The sample generation process is guided by user-defined parameters and informed by statistical insights from the entire dataset, including class distributions across each dimension. This compact representative sample is then provided to a generative LLM (e.g., GPT-4), which, guided by a tailored prompt, appropriate parameters, and context information, produces a human-readable summary that is both balanced and comprehensive, effectively highlighting a product's or service's strengths and weaknesses while ensuring fair representation of consumer feedback.

4 Experimental Results

The experimental evaluation of our framework is based on two newly created datasets, addressing the lack of publicly available AI-generated summaries. While platforms like Amazon and Tripadvisor now provide AI-generated summaries, others such as Booking and TrustPilot do not, limiting their inclusion in existing datasets[3]. To fill this gap, we collected user reviews and their AI-generated summaries for selected products and hotels from Amazon and Tripadvisor. These datasets, along with the code of main components, are publicly available at https://github.com/SCAlabUnical/UserReviewDatasets/.

The first dataset consists of Amazon product reviews, primarily in the electronics category, containing approximately 10,000 reviews across hundreds of products. It includes key attributes such as user ratings (one to five stars), review titles, descriptions, reactions, verification status, and metadata like location and date. The second dataset comprises hotel reviews from Tripadvisor, focusing on hotels in New York. It includes review ratings, titles, travel dates, and hotel details such as address, coordinates, and number of reviews. Additional attributes include subcategory ratings (e.g., value, service, location), responses from property owners, user-uploaded photos, and trip type classification (couples, solo, family, business, friends, or unspecified).

To evaluate our framework, we applied BERT-based models for multidimensional classification of sentiment, emotions, and topics, then generated structured summaries and compared them with AI-generated ones. Performance was

[3] `https://amazon-reviews-2023.github.io/main.html`

assessed using quantitative metrics (text scores, latent representations) and qualitative evaluations (automated and manual). The following sections detail the classification process (Section 4.1), summary generation (Section 4.3), and comparative performance analysis (Section 4.4).

4.1 Multi-Dimensional Classification Using BERT Models and topic extraction

As discussed in Section 3, we employ BERT-based classifiers to extend the information contained in the reviews. This multidimensional data enrichment process can significantly help generative models to produce comprehensive summaries. In particular, we trained and utilized classifiers for the following dimensions: *(i) Sentiment*, determining whether a review conveys a positive or negative sentiment; *(ii) Topic*, which associates the subject matter discussed in a review (in this case, topics cannot be defined a priori but are derived from a dedicated topic extraction process); *(iii) Emotion*, which identifies the emotional tone and expressions conveyed within the text, including anger, disgust, joy and surprise.

Careful evaluations were conducted to select the best classification models for each dimension, following approaches used in prior work [29,5]. For sentiment and emotion classification, we fine-tuned BERT-based models on annotated datasets, achieving the best performance in terms of AUC scores. For topic detection, we employed BERTopic [9], which outperformed alternative methods in both consistency and diversity of topics. Unlike sentiment and emotion analysis, topic detection was performed collectively on all reviews of a product to extract dominant themes, optimizing coherence values to determine the ideal number of topics.

4.2 Polarization-Driven Stratified Sampling for Relevant Review Selection

To calculate the most relevant reviews for analysis among dimensions of interest, we use the following method. Consider an initial dataset of reviews R, where each review $r \in R$ is associated with one or more dimensions $d_1, d_2, \ldots, d_k$. Each dimension d_i has a set of possible classes $C(d_i) = \{c_1, c_2, \ldots, c_m\}$. For each review r and each dimension d_i, there is an associated probability distribution over the classes $c \in C(d_i)$, denoted as $P(c \mid d_i)$. To create a representative sample S of N reviews for analysis:

1) Select Dimensions and Classes: the user identifies dimensions of interest $\{d_1, \ldots, d_z \mid z \geq 1\}$, which are relevant for analysis. For each selected dimension, specific classes $C'(d_i) \subseteq C(d_i)$ may also be chosen based on the scope of the analysis.

2) Compute Class Distributions: for each selected dimension d_i, calculate the probability $P(c \mid d_i)$ of reviews in R that belong to each class $c \in C'(d_i)$.

3) Allocate Sample Sizes: for each class $c \in C'(d_i)$, determine the number of reviews $N_{c|d_i}$ to include in the sample: $N_{c|d_i} = P(c \mid d_i) \cdot N$

4) Rank Reviews by Polarization: assign a confidence score to each review $r \in R$, reflecting the degree of polarization across the selected dimensions. For each dimension d_i, a statistical measure of the distribution $P(c \mid d_i)$ is used to calculate the confidence. The confidence score is designed to assign a value of 1 to *fully polarized* distributions, a value of 0 to *neutral* distributions, and intermediate values between 0 and 1 to distributions ranging from slightly to strongly polarized. Several statistical measures can be used to calculate the confidence score, including variance of probabilities, entropy, Gini impurity, and Kullback-Leibler divergence. Among these, we chose the *variance of probabilities* for its simplicity, interpretability, and ability to emphasize polarization while normalizing across dimensions. The confidence score for each dimension is calculated as:

$$\text{Confidence}(r, d_i) = 1 - \frac{\text{Var}(P(c \mid d_i))}{\text{MaxVar}(d_i)}$$

where $\text{MaxVar}(d_i) = \frac{m-1}{m^2}$ and $m = |C'(d_i)|$, the number of classes in dimension d_i. The combined confidence score for a review is computed by aggregating the confidence scores across all dimensions:

$$\text{Confidence}(r) = \frac{1}{z} \sum_{d_i} \text{Confidence}(r, d_i)$$

where z is the total number of dimensions considered.

By normalizing the variance for each dimension, this method ensures consistency across dimensions with varying numbers of classes, making it an effective and computationally efficient choice for evaluating polarization in multidimensional datasets.

5) Polarization-Based Review Sampling Using Knapsack: iterate through the ranked list of reviews, where reviews are ordered by their confidence scores. The ordering prioritizes reviews that are fully polarized across all dimensions, followed by those that are strongly polarized, slightly polarized, and finally the neutral ones. This ranking ensures that the most polarized reviews, which provide clearer signals across dimensions, are considered first for inclusion in the sample S. During this process, add reviews to the sample S, ensuring that the number of reviews for each class $N_{c|d_i}$ does not exceed the allocated target for the respective class. This approach guarantees that the sample S reflects the specified class distributions across all selected dimensions while emphasizing reviews with greater polarization for more meaningful analysis.

6) Final Adjustment: if S does not meet the exact sample size N due to rounding or constraints, adjust the sample by adding or removing reviews with the lowest confidence scores, ensuring that the class distributions remain approximately consistent.

This method ensures that the sample S is representative of the class distributions across the selected dimensions and classes, aligning with the objectives of the specific analysis. By leveraging a knapsack-inspired approach, the

sampling process prioritizes reviews with higher polarization, ensuring a balanced yet informative subset that captures the most relevant signals for analysis while adhering to predefined class distribution constraints. To illustrate how the polarization-based sampling method evaluates and prioritizes reviews, consider a dataset with two key dimensions: d_1 *(Sentiment)*, comprising two classes (*Positive* and *Negative*), and d_2 *(Topic)*, comprising three classes (*Topic1*, *Topic2*, and *Topic3*). By examining examples of fully polarized and neutral reviews, we can observe how the confidence score reflects the degree of polarization across dimensions. For instance, a review such as *((1, 0), (1, 0, 0))* demonstrates complete polarization, achieving a confidence score of *1*. Conversely, a fully neutral review like *((0.5, 0.5), (0.333, 0.333, 0.333))* exhibits no polarization, resulting in a confidence score of *0*.

From the point of view of algorithmic complexity, $O(|R| \log |R| + |R| \cdot z)$ represents the complexity of the algorithm, where $|R|$ is the number of reviews, $O(|R| \log |R|)$ accounts for the sorting step, and $O(|R| \cdot z)$ arises from the Knapsack-Based Sampling across z dimensions. This ensures the method is efficient and scalable for datasets with a high number of reviews and dimensions.

4.3 Review Summary Using Generative Models

In this phase, we leverage insights from multi-dimensional classification to create comprehensive summaries on sentiments, emotions, and topics by interacting with generative models like GPT via API. These models automate the generation of structured content, ensuring a balanced and thorough analytical perspective. In particular, we use the GPT-4o API with a temperature setting of 0 to ensure accurate and consistent outputs. Lower temperature values minimize randomness, resulting in greater consistency, while higher values introduce more variability and diversity. We have defined a prompt that is structured to guide GPT in generating human-readable summaries about a specific product (or hotel), starting from key elements such as the product name, a short description, and a curated list of reviews. Below is the prompt approach, called `GPT-adv`, that we used to summarize product reviews on Amazon:

> **Product (`$p`)**:{*monitored product*}, **Description (`$d`)**:{*product description*},
>
> **Reviews (`$R`)**: [*/*1st review*/* {*Title (`$tr`)*:{*title of the review*}, *Text (`$t`)*:{*text of the review*}, *Sentiment (`$s`)*:{*positive, neutral or negative*}, *Topics (`$t`)*:{*main topics addressed*}, *Emotions (`$e`)*:{*main emotions expressed*}}, */*2nd review*/* {...}, ...]
>
> **Input**: A list of reviews `$R` for the product `$p` described in `$d`. Each review includes its title, full text, a sentiment label, the main topics addressed, and the primary emotions expressed.
>
> **Task**: Generate a comprehensive and balanced report about the product that captures the essence of all the reviews by summarizing their key points and covering all significant aspects, while remaining concise. The report must be a single paragraph without line breaks or colons and should not exceed `$N` words.

Specifically, the `GPT-adv` prompt receives a list of reviews (`$R`) via API. Notably, this list is generated using a stratified sampler to ensure a balanced

set of reviews that covers all the dimensions considered, including the classes of each dimension. Each provided review is derived from a multidimensional classification and includes the title and text of the review, a sentiment score, the identified emotions, and the topic represented. The goal is to produce a summary that highlights key strengths and weaknesses while adhering to a specified word count limit (`$N`).

We evaluate the outputs of `GPT-adv` against those generated by AI tools on online platforms (such as Amazon and Tripadvisor) and a baseline approach that processes only the original reviews via a chat interface (i.e., without leveraging additional classification), referred to as `GPT-base`. This comparison illustrates that selecting a balanced sample of reviews not only significantly enhances the quality of the generated summary by providing a more representative and nuanced view, but also reduces token usage by condensing the input to the most informative subset, thereby ensuring that a good number of tokens is used efficiently.

When using ChatGPT-4o via API, the input limit is 128,000 tokens (approximately 96,000 words or 6,400 tweets, assuming 15 words per tweet). While file uploads of up to 512 MB are allowed, only the portion that fits within the context window is processed; content exceeding this limit remains unprocessed. It is important to note that as the token count approaches the maximum capacity of the model, performance may decrease, especially for tasks involving long and complex content [27].

4.4 Performance evaluation

In this section, we describe the performance evaluation of our framework against basic approaches and AI-generated summaries available on platforms such as Amazon and Tripadvisor. As previously outlined, our goal is to evaluate how the extra data provided to GPT via multidimensional classification (sentiment, extracted topic, and detected emotion) helps develop more precise commentaries. By leveraging this additional input, our approach delves deeply into the subject and uncovers subtle effects of various information inputs on GPT's ability to generate commentaries. Notably, our framework also aims to define a compact sample that minimizes token usage while still producing comprehensive summaries that capture all aspects of the reviews, ensuring efficient processing without sacrificing detail.

4.5 Step-by-Step Operation and Parameter Configuration

For demonstration purposes, we illustrate our framework using an electric toothbrush from Amazon (anonymized despite being a specific model) as a case study. Below, we present the product along with examples of both a positive and a negative review. For each review, our framework identifies the associated sentiment, referenced topic, and detected emotion. These examples highlight how our system classifies user feedback into multidimensional categories, establishing the

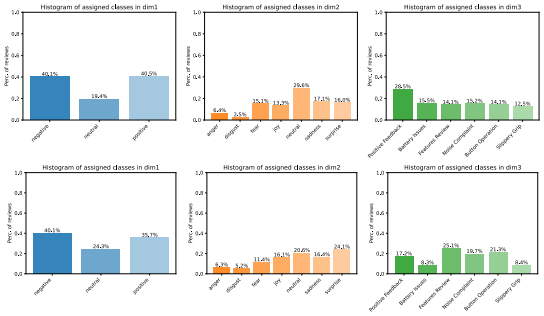
basis for generating a compact, balanced sample that captures comprehensive insights for subsequent summarization.

`Product (p)` = "*[Anonimized]* Rechargeable Electric Toothbrush", `Description (d)` = "Protect your gums with sensi cleaning mode and gum pressure control..."

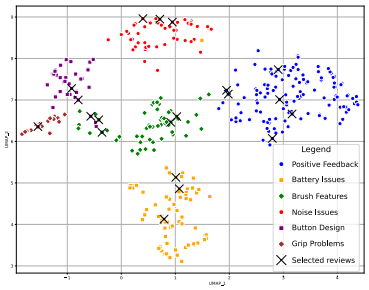
`Example of positive review (pr)` = {*title (tr)* : "Works Well.", *text (t)*:"... When I purchased this, I was half expecting to send it back. I was pleasantly surprised by how well this one works and cleans my teeth...", *sentiment (s)*: `Positive`, *topics (t)*: `Positive feedback`, *emotions (e)*: `Surprise`}

`Example of negative review (nr)` = {*title (tr)*: "Very Slippery. Hard to find on/off switch.", *text (t)*: "I needed to replace my *[anonimized]* electric toothbrush. Made a big mistake buying this one. The handle is too slippery, and the onoff switch looks nothing like the picture. A small button that is hard to find. Will be replacing as soon as I find another one.", *sentiment (s)*: `Negative`, *topics (t)*: `Button Design`, `Grip problems`, *emotions (e)*: `Sadness`}

Subsequently, our knapsack-based selection method preserves the class distributions across all dimensions when reducing the full dataset to a compact sample of `N` reviews (here, `N`=20). As shown in Figure 2(a), the percentage distributions of sentiment, topics, and emotion in the sample closely mirror those of the full dataset. Figure 2(b) further confirms that the 20 selected reviews—marked with an "X"—are evenly distributed among the six topic clusters extracted using BERTopic and compressed via UMAP, demonstrating the effectiveness of our balanced selection process. Figure 3(a) shows the distribution and density of sentiment classes (only positive and negative classes) for both the full set and the selected sample. The chart demonstrates that the sentiment distribution in the compact sample closely mirrors that of the full dataset, and similar balancing was achieved for the other dimensions.



(a) The top chart shows the full dataset's class distribution (sentiment, topics, emotion), while the bottom chart shows the sample.



(b) Clustered reviews with topics extracted by BERTopic and visualized via UMAP. "X" marks denote sample reviews.

Fig. 2. Class distributions for full dataset and balanced sample (N=20).

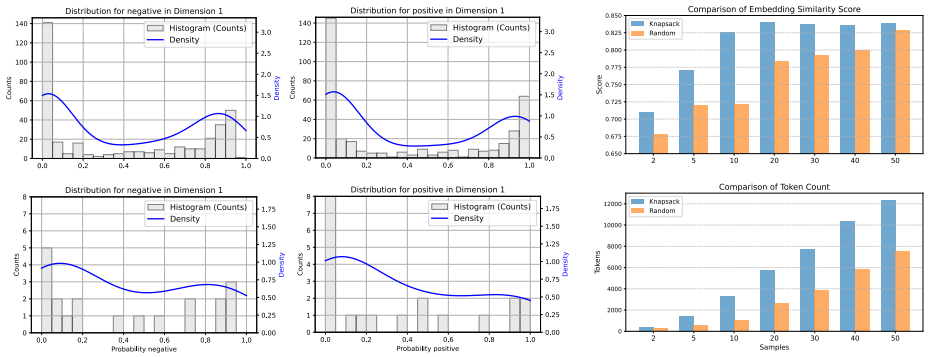
To verify that a small but representative sample can generate a quality summary, we first generated a full summary using all available reviews retrieved via the API for a product (about 500) and then produced additional summaries using samples of `N` reviews. We compared the embeddings of these sample-based summaries with the full-dataset summary using cosine similarity, averaging the results over 10 different products. Figure 3(b) (at the top) illustrates these comparisons and contrasts our balanced selection approach with random sampling. Using ModernBERT [25] to compute the text embeddings, our method achieves a cosine similarity of 0.85 with a sample of 20 reviews, indicating that the embeddings share high semantic similarity and many common features [16]. Notably, increasing the sample size beyond 20 does not significantly improve this value. Cosine similarity rarely reaches 1 because even summaries that capture the same information can differ in wording, sentence order, and subtle nuances, resulting in slight variations in their embeddings. In contrast, the random selection approach gradually improves with increasing `N`, but only matches our balanced strategy when the sample size exceeds 50 reviews.

Additionally, as illustrated in Figure 3(b) (bottom), the token count required for summarization increases substantially with the number of reviews considered. Our balanced selection approach uses only about 5,000 tokens for 20 reviews, whereas summarizing all reviews (i.e. 500) requires over 50,000 tokens—a nearly tenfold increase. This marked reduction boosts computational efficiency and lowers costs associated with large language models, all without sacrificing summary quality. Notably, our method tends to select longer, information-dense reviews, which may result in a slightly higher token count per review compared to random sampling; however, the overall token usage remains dramatically lower than processing the entire dataset. Therefore, for our experiments on products and hotels, we set the sample size to 20 reviews—roughly 5,000 tokens—to achieve an optimal balance between quality and efficiency.

In the final step, we compare the commentaries generated by three distinct approaches for our reference product (electric toothbrush). First, we show the commentary produced by the Amazon AI system, the summary generated by `GPT-base`, which uses a file containing all reviews and a prompt, and `GPT-adv` the commentary produced by our framework.

Amazon AI-generated = Customers like the cleanability, performance, value, and quality of the toothbrush. For example, they mention that it effectively cleans between teeth, removes up to 100% more plaque along the gumline, and has a good price point. Customers also appreciate ease of use. That said, opinions are mixed on durability, and battery life.

GPT-base = User reviews indicate that this electric toothbrush excels in design and cleaning performance. Many users note its comfortable fit, which allows the brush to reach difficult areas such as the molars, and its built-in timer and pressure sensor enhance the brushing experience. Its robust battery life and user-friendly interface ensure consistent performance over time, and the design effectively minimizes plaque buildup along the gumline, contributing to improved oral hygiene. However, a few users have reported concerns about its durability, with some experiencing wear on the handle after prolonged use. Despite these durability issues, the overall balance between advanced features and



(a) The top chart shows the counts and density of sentiment instances (positive/negative) for the full dataset, while the bottom chart shows these for the balanced sample.

(b) Summary embeddings and token counts for full dataset vs. sample using our and random selection.

Fig. 3. Class distributions (full dataset vs. balanced sample, N=20) and summary embedding/token usage: balanced vs. random.

affordability makes this toothbrush a popular choice among consumers seeking a reliable, efficient dental care solution.

GPT-adv = The [...] electric toothbrush has garnered mixed reviews from users, highlighting both its strengths and weaknesses. Many users appreciate its effectiveness in plaque removal and the thorough clean it provides compared to manual brushes, with some noting significant improvements in oral hygiene. The timer feature, which prompts users to switch quadrants every 30 seconds, is also seen as a helpful addition. However, several reviewers express frustration with the charging system, citing long charging times and poor battery life, which detracts from the overall experience. Design issues, particularly with the on/off button being difficult to press and the handle lacking grip, have been noted as significant drawbacks, especially for those with limited hand strength. Additionally, some users report that the brush can be quite loud and that the replacement heads can be costly. Overall, while the toothbrush performs well in cleaning, its usability and design flaws may lead some to seek alternatives.

Comparing the outputs, the Amazon AI-generated and `GPT-base` reviews emphasize positive aspects—highlighting features like cleanability, performance, and design—while only briefly mentioning minor issues. In contrast, `GPT-adv` offers a more comprehensive analysis by addressing both the product's strong cleaning performance and its critical drawbacks, such as poor battery life and a slippery handle. This advanced method not only reduces token usage dramatically while preserving the overall review characteristics, but also delivers a more balanced commentary that captures both strengths and weaknesses in detail.

4.6 Aggregate Metrics and Comparative Analysis

Here, we present a comprehensive analysis conducted to compare the quality of generated summaries against the original text. First, we assessed the quality

of the synthesis using established semantic metrics, such as TF-IDF, ROUGE, S-BERT, S-RoBERTa, BERTScore, and BLANC. Second, we employed ChatGPT as an evaluator to assign scores based on various aspects, including topic coverage, clarity, and readability. Finally, we carried out a human evaluation, who assessed the generated summaries through a survey.

Analysis of Semantic Metrics To evaluate the quality of these reports, we used a set of commonly applied metrics for assessing summary quality against a reference text. In the absence of a specific reference text, we defined the reference as the concatenation of all reviews describing a product or a hotel. Then, the following metrics were considered: *i*) TF-IDF for lexical similarity; *ii*) ROUGE-1 for measuring unigram overlap and basic content recall; *iii*) ROUGE-2 for evaluating bigram overlap and capturing short-sequence coherence; *iv*) S-BERT and *v*) S-RoBERTa for sentence-level embeddings to assess deeper contextual understanding; *vi*) BERT-Score for evaluating fine-grained semantic similarity at the word level; and *vii*) BLANC-help for assessing fluency and informativeness.

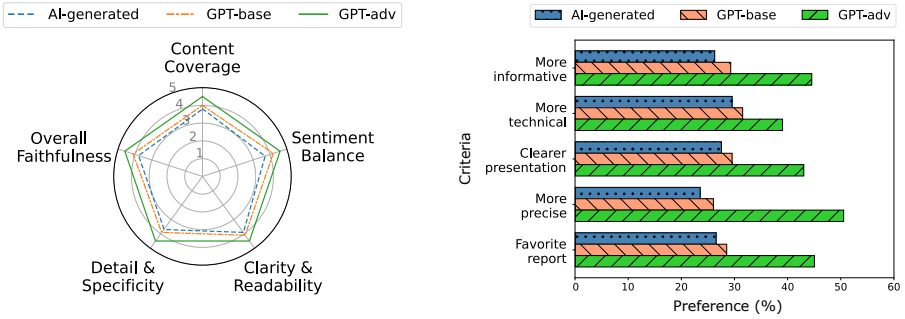
Table 1. Evaluation of semantic scores in the Amazon and Tripadvisor case studies for different approaches.

	TF-IDF	Rouge-1	Rouge-2	S-BERT	S-RoBERTa	Bert-Score	BLANC-help
AI-generated	0.237	0.016	0.002	0.565	0.800	0.520	0.034
GPT-base	0.253	0.016	0.004	0.614	0.798	0.509	0.034
GPT-adv	**0.294**	**0.019**	**0.007**	**0.623**	**0.813**	**0.568**	**0.054**

Table 1 reports the average values of metrics obtained by summaries generated using different approaches (`AI-generated`, `GPT-base`, `GPT-adv`) across the two case studies (Amazon and Tripadvisor). In both case studies, the evaluation reveals a clear improvement when moving from simpler approaches, such as `AI-generated` and `GPT-base`, to more advanced ones like `GPT-adv`. A higher *TF-IDF* score reflects an enhanced ability to capture and synthesize the core content of the reviews. Additionally, higher *ROUGE-1* and *ROUGE-2* values indicate that the summary conveys the details of the original reviews more effectively, sharing similar phrasing and structure. Metrics like *S-BERT*, *S-RoBERTa*, and *BERTScore* further demonstrate that the advanced approach better captures semantic similarities. Finally, the improvements in *BLANC-help* highlight superior contextual flow and coherence in the summaries, making them clearer and more comprehensive.

ChatGPT evaluation This section presents a detailed evaluation of the summaries generated for both case studies using ChatGPT as the evaluator [13]. We asked ChatGPT to assess the following five dimensions, where each is scored on a 5-point scale with higher values indicating better performance: *i*) *Content Coverage*, which evaluates how well the summary captures key aspects from the

reviews (e.g., position and cleaning for an hotel or design and battery life for a smartphone); *ii*) *Sentiment Balance*, measuring whether the summary proportionally reflects both positive and negative feedback; *iii*) *Clarity & Readability*, which assesses ease of understanding, well-structuring and clarity; *iv*) *Detail & Specificity*, indicating how specific the summary is regarding individual features (e.g., "removes up to 100% more plaque" or "built-in timer and pressure sensor"); and *v*) *Overall Faithfulness*, which verifies accuracy and alignment with the original reviews without distortion or exaggeration.



(a) ChatGPT evaluation. (b) Human-Made evaluation.

Fig. 4. Evaluation of the reports by ChatGPT and humans on the two considered case studies (average values).

Figure 4(a) illustrates the average evaluation results for the two case studies, highlighting that while both the `AI-generated` and `GPT-base` summaries focus largely on positive aspects, the `GPT-adv` approach provides a more balanced and comprehensive summary by effectively capturing both strengths and weaknesses.

Human-Made Evaluation Regarding the human evaluation, we conducted a survey involving 20 participants and 10 products/hotels. Each participant was presented with a set of reviews for these products and asked to rate the summaries generated by the three considered approaches (`AI-generated`, `GPT-base`, and `GPT-adv`) without being informed about the approach used to generate each summary. To ensure unbiased evaluations, the different versions were presented in a randomized order. Specifically, they were asked to answer these questions: *i*) which summary provides more overall information content? *ii*) which summary includes more technical or specialized aspects? *iii*) which summary offers a clearer presentation? *iv*) which summary demonstrates greater precision and clarity in its contents? *v*) which summary do you prefer for overall quality?

Figure 4(b) shows the percentage of participants who preferred the base, advanced, and AI-generated summaries across the five criteria considered. As shown, participants consistently favored `GPT-adv` over other versions across all

aspects, though the preference for `GPT-adv` was only slightly higher than for `GPT-base`. In particular, `GPT-adv` received higher ratings, particularly for its greater information content and clearer presentation. Notably, evaluators found summaries generated by `GPT-base` slightly better than those by `AI-generated`, indicating that the latter tends to produce summaries that are less balanced and often omit important details.

5 Conclusions

User-generated reviews play a crucial role in shaping business strategies, and LLMs like BERT and GPT have significantly improved their analysis. This paper introduced a novel framework that enhances user review summarization by leveraging LLMs to ensure balanced and comprehensive insights. Unlike conventional AI-generated summaries, which often exhibit positive bias, our approach systematically classifies reviews across multiple dimensions, such as sentiment, emotion, and topic, before applying a stratified sampling method to create a representative subset. By incorporating a knapsack-based selection strategy, we effectively balance review content while optimizing token usage, leading to high-quality summaries with significantly reduced computational cost. An extensive evaluation on Amazon and Tripadvisor datasets has been carried out, using both quantitative and qualitative measures, demonstrating that our approach outperforms existing summarization techniques, including those employed by major online platforms. Future work includes extending the framework to compare similar products, applying it to diverse datasets (e.g., TrustPilot, Yelp, Google Reviews, Reddit), and analyzing opinions on social media pages to uncover strengths, weaknesses, and areas for improvement. Additionally, the proposed approach could also be used to analyze opinions on the social media pages of political figures, institutions, or companies, summarizing strengths, weaknesses, user requests, and areas of engagement for better understanding and potential improvements.

Acknowledgments. 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).

References

1. Aakash, Gupta, S., Noliya, A.: Url-based sentiment analysis of product reviews using lstm and gru. Procedia Computer Science **235**, 1814–1823 (2024). `https://doi.org/https://doi.org/10.1016/j.procs.2024.04.172`, `https://www.sciencedirect.com/science/article/pii/S1877050924008482`, int. Conf. on Machine Learning and Data Engineering (ICMLDE 2023)
2. Anbumani, P., Selvaraj, K.: Enhancing sentiment analysis classification for amazon product reviews using cnn-sigtan-beta activation function. Multimedia Tools and Applications **83**(19), 56719–56736 (2024)

3. Ayanouz, S., Abdelhakim, B.A., Benhmed, M.: A smart chatbot architecture based nlp and machine learning for health care assistance. In: 3rd Int. Conf. on Networking, Information Systems & Security (2020)
4. Caldarini, G., Jaf, S., McGarry, K.: A literature survey of recent advances in chatbots. Information **13**(1) (2022)
5. Cantini, R., Cosentino, C., Marozzo, F.: Multi-dimensional classification on social media data for detailed reporting with large language models. In: IFIP Int. Conf. on Artificial Intelligence Applications and Innovations. pp. 100–114. Springer (2024)
6. Cantini, R., Cosentino, C., Marozzo, F., Talia, D., Trunfio, P.: Harnessing prompt-based large language models for disaster monitoring and automated reporting from social media feedback. Online Social Networks and Media **45**, 100295 (2025)
7. Cosentino, C., Gunduz-Cure, M., Marozzo, F., Ozturk-Birim, S.: Exploiting large language models for enhanced review classification explanations through interpretable and multidimensional analysis. In: 27th Int. Conf. on Discovery Science (DS2024) (2024)
8. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. National Academy of Sciences **120**(30) (2023)
9. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv:2203.05794 (2022)
10. Hossain, M., Habib, M., Hassan, M., Soroni, F., Khan, M.M.: Research and development of an e-commerce with sales chatbot. In: 2022 IEEE World AI IoT Congress (AIIoT). pp. 557–564 (2022)
11. Li, H., Zhang, Y., Zhang, R., Chaturvedi, S.: Coverage-based fairness in multi-document summarization (2025), https://arxiv.org/abs/2412.08795
12. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: 14th Int. Conf. on World Wide Web. pp. 342–351 (2005)
13. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023)
14. Mayfield, J.e.a.: On the evaluation of machine-generated reports. In: 47th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. p. 1904–1915. SIGIR '24, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657846, https://doi.org/10.1145/3626772.3657846
15. Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Andia, M., Tejos, C., Prieto, C., Capurro, D.: A survey on deep learning and explainability for automatic report generation from medical images. ACM Computing Surveys (CSUR) **54**(10s), 1–40 (2022)
16. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. Computational Linguistics **32**(1), 1–24 (2006)
17. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence **2**, 100033 (2021)
18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. cs/0205070 (2002)
19. Perti, A., Sinha, A., Vidyarthi, A.: Cognitive hybrid deep learning-based multimodal sentiment analysis for online product reviews. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **23**(8) (aug 2024). https://doi.org/10.1145/3615356, https://doi.org/10.1145/3615356
20. Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. Natural Language Processing Journal **6**, 100056 (2024)

21. Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. Natural Language Processing Journal **6**, 100056 (2024). https://doi.org/https://doi.org/10.1016/j.nlp.2024.100056, https://www.sciencedirect.com/science/article/pii/S2949719124000049
22. van Schaik, T.A., Pugh, B.: A field guide to automatic evaluation of llm-generated summaries. In: 47th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. p. 2832–2836. SIGIR '24, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3661346, https://doi.org/10.1145/3626772.3661346
23. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G.: Text classification via large language models. arXiv:2305.08377 (2023)
24. Wang, F., Xu, Z., Szekely, P., Chen, M.: Robust (controlled) table-to-text generation with structure-aware equivariance learning. arXiv:2205.03972 (2022)
25. Warner, B.e.a.: Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663 (2024)
26. Yang, B., Liu, Y., Liang, Y., Tang, M.: Exploiting user experience from online customer reviews for product design. Int. Journal of Information Management **46**, 173–186 (2019)
27. Yuan, T., Ning, X., Zhou, D., Yang, Z., Li, S., Zhuang, M., Tan, Z., Yao, Z., Lin, D., Li, B., et al.: Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. arXiv preprint arXiv:2402.05136 (2024)
28. Zaremba, A., Demir, E.: Chatgpt: Unlocking the future of nlp in finance. Modern Finance **1**(1), 93–98 (2023)
29. Zhang, H., Shafiq, M.O.: Survey of transformers and towards ensemble learning using transformers for natural language processing. Journal of big Data **11**(1), 25 (2024)
30. Zhang, K.Z., Zhao, S.J., Cheung, C.M., Lee, M.K.: Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. Decision support systems **67**, 78–89 (2014)
31. Zhong, A., Mo, D., Liu, G., Liu, J., Lu, Q., Zhou, Q., Wu, J., Li, Q., Wen, Q.: Logparser-llm: Advancing efficient log parsing with large language models. In: 30th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. p. 4559–4570. KDD '24, New York, NY, USA (2024). https://doi.org/10.1145/3637528.3671810, https://doi.org/10.1145/3637528.3671810