# AI-Driven Sentiment Analysis: A Unified Framework for Strategic Insights in Tourism

Nikolaos Gkaripis *, Georgios Trichopoulos, George Caridakis

Posted Date: 15 January 2026

doi: 10.20944/preprints202601.1176.v1

Keywords: artificial intelligence; sentiment analysis; cultural heritage; large language models; tourism; hospitality; recommendation systems

 doi:10.20944/preprints202601.1176.v1



*Article*

# AI-Driven Sentiment Analysis: A Unified Framework for Strategic Insights in Tourism

**Nikolaos Gkaripis, Georgios Trichopoulos and George Caridakis**

Department of Cultural Technology and Communication, University of the Aegean, 81100, University Hill, Mytilene, Greece
* Correspondence: nikosgkaripis@iti.gr

**Abstract**

This paper presents an Artificial Intelligence (AI) -driven framework designed to bridge the gap between raw user feedback and strategic decision-making. Moving beyond traditional sentiment analysis, which often overlooks the specific "why" behind visitor dissatisfaction, this research utilizes a sophisticated dual approach. By integrating the contextual precision of Bidirectional Encoder Representations from Transformers (BERT) with the generative reasoning of Large Language Models (LLMs) like Gemini, the system extracts fine-grained, aspect-based insights and actionable recommendations. The framework's effectiveness is demonstrated through a case study of the Archaeological Site of Mystras. Ultimately, this work offers a scalable solution for tourism professionals and policymakers to listen more deeply to the authentic voice of the traveler.

**Keywords:** artificial intelligence; sentiment analysis; cultural heritage; large language models; tourism; hospitality; recommendation systems

## 1. Introduction

The hospitality and tourism industries have been profoundly transformed by digital technologies, completely changing how people plan their trips, share their experiences, and evaluate destinations. In today's era of Big Data and AI, User-Generated Content (UGC) such as social media posts, online reviews, and ratings on platforms like TripAdvisor and Google Maps plays a crucial role in the tourism industry. This content has become one of the most trusted and influential sources of information, shaping travelers decisions while also providing valuable insights for tourism professionals and not only [1]. Instead of relying on well structured advertising campaigns or commercials, travelers increasingly look to the authentic opinions and lived experiences of other users as a more reliable option feeling they ask a friend and not a company. This shift has made big data analytics a crucial tool for understanding traveler behavior, expectations, and emerging trends.

As Kaplan and Haenlein observed [2], the ease with which users can create and share content has democratized information, making it easy and accessible , giving consumers a powerful voice and reducing the control once held by organizations. In tourism, this means perceptions can change almost overnight. Travelers are no longer passive recipients of services. They actively shape the value of destinations through reviews, photos, and comments as the have more options to evaluate a destination. Even a single post can influence how a hotel, attraction, or entire city is perceived, pushing tourism operators to adopt more responsive and forward-thinking digital strategies to manage their online presence and reputation as they fear of losing their customers.

At the same time, this constant stream of online content presents significant challenges. Tourism businesses, cultural institutions, and public authorities are often overwhelmed by thousands of unstructured reviews, making manual analysis both time-consuming and costly without effective preprocessing techniques [3]. Much of these data are messy and filled with repetition, informal language, and inconsistent formatting, making meaningful interpretation difficult. To deal with this

 doi:10.20944/preprints202601.1176.v1



complexity, automated solutions have become increasingly necessary, with Sentiment Analysis (SA) standing out as a practical way to turn large volumes of raw text into clear, actionable insights [4].

While web scraping and Application Programming Interfaces (API) are effective for gathering user data from online platforms [5], uncovering its real value requires more advanced Natural Language Processing (NLP) techniques. These tools must go beyond simple word matching and learn to interpret the nuances of human communication, including cultural differences, personal bias, sarcasm, irony, and the ever-changing language of the internet.

In bibliography, SA has focused mainly on polarity classification — simply determining if a text is positive, negative, or neutral [6] [7]. Although this broad approach is useful for getting a general sense of customer satisfaction, it often lacks the detail needed to drive real operational improvements in a world in which data growing rapidly small but important details can easily lost [8]. For instance, a hotel review might receive a high 4-star rating, effectively masking serious underlying complaints about cleanliness or staff behavior. This kind of specific feedback often gets lost in the overall score, making it difficult for businesses to know exactly what to fix [9].

To address this limitation, the field has shifted toward Aspect-Based Sentiment Analysis (ABSA). Unlike general SA, ABSA breaks a review down into specific "aspects" such as "accommodation", "service" or "cultural experience" and assigns a sentiment score to each one individually [6]. This level of detail allows stakeholders to examine exactly what drives customer satisfaction or dissatisfaction, enabling them to make targeted improvements rather than generic responses that would not benefit them in a long term [10]. However, building reliable ABSA systems remains a significant technical challenge. It requires the model to do more than just identify keywords; it must correctly link sentiments to the right aspects — a task that becomes notoriously difficult when dealing with irony, sarcasm, or implied emotions, which still confuse even advanced deep learning models [11].

The world of NLP changed forever with the arrival of Transformer architectures, such as BERT. Before this, models read text sequentially, which often limited their understanding of complex context making them unable to generalize and nowhere near human behavior. Transformers broke this barrier by analyzing entire sequences at once, capturing the deep, two-way context of words and their relationships [12]. This technological leap set the stage and prepared the ground for the current era of LLMs, including powerful systems like the Generative Pre-trained Transformers (GPT) series (GPT-3, GPT-4) and Google's Gemini. Unlike their predecessors, which were often built for one specific job, these massive models represent a shift toward general-purpose intelligence, capable of handling a vast array of linguistic tasks right out of the box and not only, make them able to create content like images or videos from scratch approach human like intelligence.

What truly sets these modern generative models apart is their ability to learn and reason in ways that feel almost human. Trained on enormous libraries of diverse text, they possess "few-shot" and "zero-shot" learning capabilities — meaning they can perform complex tasks they have not explicitly been trained for, often with little to no instruction [13,14]. Instead of just categorizing text based on keywords, they function as versatile cognitive engines that can draft emails, write code, or summarize dense reports. Comprehensive surveys of the GPT family highlight this versatility, noting that these models function less like simple classification tools and more like adaptable reasoning agents that can generalize their knowledge across completely different domains creating a variety of content [15].

However, using these powerful LLMs in specialized high-stakes fields such as tourism management is not as simple as it seems. Although LLMs are incredible at generating fluent, human-like text, rigorous research, Zhang et al. [16] points out a "performance dichotomy": they often struggle with rigid, structured tasks — like ABSA where precision matters more than creativity. In direct comparisons, smaller fine-tuned language models often outperform them in extracting specific details but there are also reliability issues. LLMs can be notoriously sensitive to "prompt engineering" where changing a single word in the question can lead to a completely different answer [17]. Furthermore, because these models are trained to be safe and polite (a process known as Reinforcement Learning from Human Feedback), they sometimes "sanitize" negative feedback, weakening the harsh but valuable

 doi:10.20944/preprints202601.1176.v1



criticisms that businesses need to hear in order to improve their services [16]. This makes it critical to systematically test these flashy new tools against traditional, reliable models to see which truly performs better in the messy, real-world environment of customer feedback in which data growing exponential every day [18].

The remainder of this research is structured as follows: Section 2 provides a comprehensive overview of Related Work, synthesizing existing literature on the evolution of SA from traditional deep learning approaches to the application of LLMs in the tourism sector. Section 3 details the Methodology, presenting the proposed end-to-end system architecture, including the data collection pipeline, the integration of the Gemini and BERT models, and the technical implementation of the visualization dashboard. Section 4 presents the Results, demonstrating the system's performance through a case study of the Archaeological Site of Mystras, supported by comparative metrics and visual analysis. Finally, Section 5 concludes with a Discussion and Future Work, interpreting the findings, addressing current limitations such as prompt sensitivity, and outlining strategic directions for future enhancements in AI-driven cultural analytics in tourism context and not only.

## 2. Related Work

SA has long been a core task in NLP, essentially teaching computers how to read human emotions and opinions hidden within text [4]. The field has changed dramatically over time. It started with simple rule-based systems that relied on static "dictionaries" of good and bad words, then moved to statistical tools like Naïve Bayes, and finally evolved into complex deep learning systems capable of learning patterns on their own [19]. A huge turning point came with the introduction of the Transformer architecture, as mentioned in Section 1. Models like BERT could finally understand the full context of a sentence, picking up on long-distance connections and subtle nuances that older models simply missed [12].

Recently, everything shifted again with the rise of LLMs like GPT-3, GPT-4 [13,14]. Unlike traditional models that needed thousands of labeled examples to learn a specific task, these giants are trained on the vastness of the internet. This gives them the incredible ability to perform new tasks with little to no extra instruction — a capability known as "zero-shot" or "few-shot" learning [13,15]. This evolution has effectively democratized AI, allowing almost anyone to analyze complex text without needing a background in data science. However, this power comes with a catch. These rapidly changing tools bring new challenges regarding their consistency and reliability. In specialized fields where precision matters, uncritical reliance on these systems should be avoided as [16,17]. One of the most critical debates in the field right now centers on a simple question: should we rely on massive, general-purpose models (LLMs), or are we better off using smaller, highly specialized ones Small Language Models (SLMs)? Zhang and his team decided to settle this with an extensive "reality check" pitting these two types of AI against each other across 26 different datasets [16].

What they found was a distinct "split personality" in the results. On one hand, generative models like ChatGPT are fantastic at the broad strokes as they can easily tell if a movie review is generally positive or negative, often matching or even beating fine-tuned models on standard tests like Internet Movie Database (IMDb) [16]. However, their performance are worse when the tasks get more structured and is there where the SLMs evaluate better.

Specifically, in ABSA — where the model needs to pinpoint exactly which feature (like "cleanliness" or "price") is being praised or criticized, the big LLMs performed approximately 50% worse than the smaller, focused models [16]. Furthermore, these large models can be surprisingly fragile; a minor change in the phrase of the prompt can lead to a completely different answer. This inconsistency makes them risky for standardized reporting. To bring some order to this chaos and suggest a solution, the authors proposed SentiEval, a unified benchmark to ensure fair comparisons. Ultimately, their findings point toward a hybrid future: one that combines the deep reasoning of LLMs with the reliable precision of SLMs that exceed in specific tasks [16].

 doi:10.20944/preprints202601.1176.v1



Mughal and his colleagues (2024) took a deeper look into the specific challenge of ABSA, pitting traditional deep neural networks against the newer wave of generative models. They specifically wanted to tackle the problem of "domain specificity" - essentially, the struggle of getting accurate results when massive amount of labeled data are missing for training [18]. By benchmarking heavy-weights like Decoding-enhanced BERT with Disentangled Attention (DeBERTa), Pathways Language Model (PaLM), and GPT-3.5-Turbo against each other, they demonstrated that the specialized tools (particularly DeBERTa) are still the champions. These fine-tuned models consistently outperformed the massive generalist LLMs when it came to the precise, surgical work of extracting specific aspect-sentiment pairs [18]. However, the gap is narrowing. The team noted that some LLMs, such as PaLM, are becoming increasingly competitive in Aspect Term Sentiment Analysis tasks. The critical takeaway here is that "one size does not fit all". A model that excels at analyzing simple product reviews might stumble when faced with the complex, cultural nuances of heritage tourism. It becomes clear that architectural selection relies on the desired outcome, a general sense of polarity versus the fine-grained accuracy of aspect extraction [18].

When the approach is related to high-stakes worlds like finance or healthcare, picking the right model gets even trickier. Fatemi and Hu [20] dove into this by investigating financial sentiment analysis, specifically asking whether it is better to fine-tune a "small" model or simply show a massive LLM a few examples (in-context learning) . Their results were surprising: they found that smaller models (ranging from 250 million to 3 billion parameters) could actually match the performance of giant, state-of-the-art LLMs, provided they were properly fine-tuned on domain-specific data [19]. Even more interestingly, while showing an LLM examples definitely helps, just throwing more examples at it does not guarantee it will get smarter. As result, when the vocabulary has special words, there is often no substitute for proper fine-tuning [20].

This observation is particularly relevant for multilingual contexts, such as the Greek market, where data is often scarcer. For example, a study on Greek medical dialogues [21] compared the generative capabilities of GPT-2 against the bidirectional context of BERT. The study found that Bert's ability to analyze context proved superior. It outperformed the generative model in accurately identifying sentiments, largely due to its ability to read context in both directions at once rather than just predicting the next word. On the flip side, however, if the goal is not just to classify text but to explain why a review is negative, generative models hold a distinct advantage. As broader surveys of the GPT-3 family suggest, their massive internet-scale pre-training allows them to reason and generate explanations in a way that specialized models simply cannot [15].

The existing literature indicates a clear trade-off: SLMs (like BERT) generally offer superior consistency and structure for classification tasks, while LLMs (like Gemini) excel in reasoning, summarization, and zero-shot adaptability. However, there is a distinct lack of accessible, deployed frameworks that allow researchers and stakeholders to directly benchmark these competing paradigms against each other in real-world cultural scenarios like reviews from well known platforms of the internet that most people use to make a critic or review of the place of interest.

This research addresses this gap by introducing a comparative analysis pipeline specifically tailored for the tourism sector. Rather than obscuring the differences by merging the models immediately, our system implements a dual-stream architecture that processes Google Maps data through both a generative LLM and a supervised SLM simultaneously. This "face-to-face" evaluation provides a transparent testbed for assessing how each architecture handles the nuances of sentiment analysis and recommendation extraction. While the current implementation focuses on benchmarking these specific models, the framework is designed as an extensible foundation, creating the way for future research to evaluate emerging architectures or develop advanced hybrid ensembles.

 doi:10.20944/preprints202601.1176.v1



## 3. Methodology

### *3.1. Pipeline Overview*

This research implements a complete, end-to-end pipeline designed to automate the extraction of sentiment and recommendations from UGC in the cultural and tourism domain mostly focusing on the reviews on google maps platform. The system architecture is structured around a flexible four-phase workflow (Figure 1), specifically engineered to support both individual model assessment and direct comparative benchmarking:

1. Data Collection & Preprocessing: User reviews are programmatically extracted and cautious cleaned to ensure high-quality input as much as possible.
2. Configurable Analysis Engine: Unlike rigid pipelines, this phase offers two distinct execution modes (Figure 2):
   - Single-Model Analysis: Executes inference using a specific architecture (either the generative Gemini or the supervised BERT) for focused evaluation.
   - All-Model Analysis (Comparative): Triggers a parallel execution of both models against the same dataset. This mode visualizes the divergence in sentiment scoring side-by-side and leverages the LLM to generate a "Combined Summary" synthesizing a holistic recommendation from the aggregated outputs.
3. Middleware Integration: A FastAPI workflow engine orchestrates these requests, routing data to the appropriate Python modules and managing the interaction between the scraping and inference layers.
4. Visualization: The processed insights are rendered via a reactive Angular v20 dashboard, which provides interactive charts for real-time model comparison.

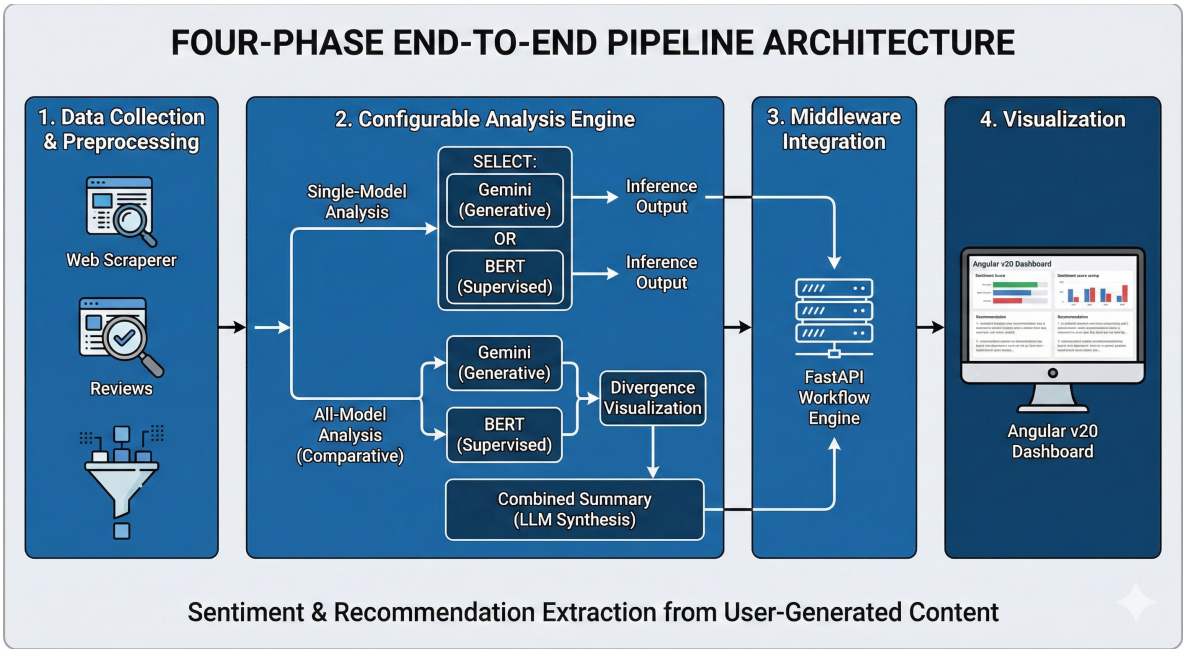


**Figure 1.** Architecture

 doi:10.20944/preprints202601.1176.v1



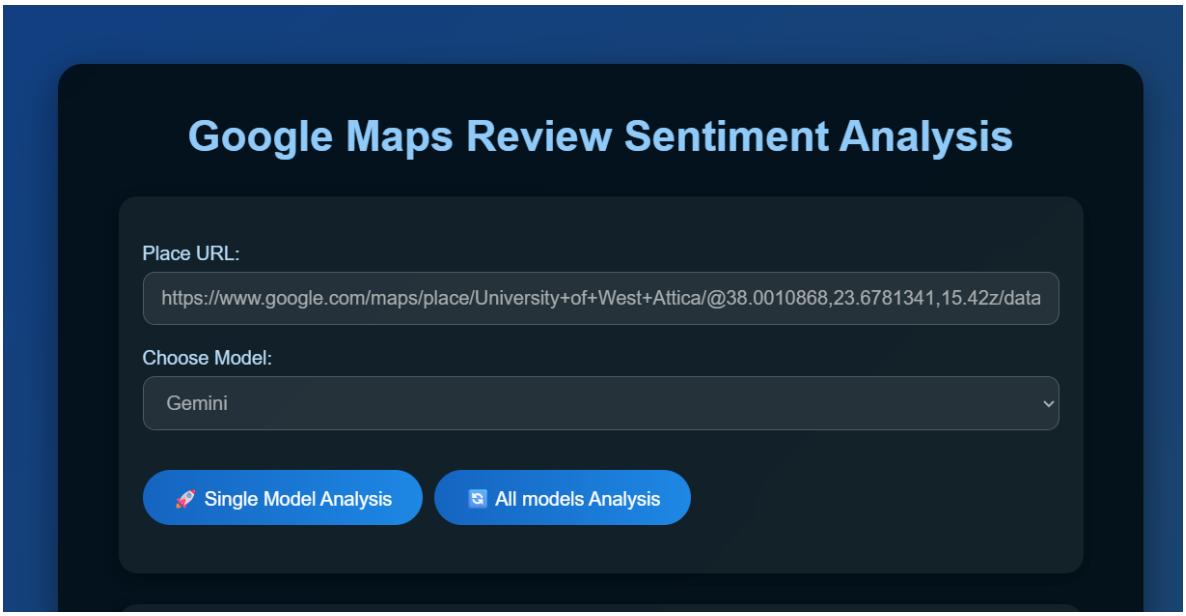


**Figure 2.** Execution modes

*3.2. Phase 1: Data Collection and Preprocessing*

The analysis begins with the acquisition of user feedback from Google Maps, selected for its authenticity and high volume of visitor interactions. The system accepts a target URL for a cultural or tourism site and programmatically extracts the available review corpus, capturing both the textual content and the associated numerical ratings.

Once the raw data are extracted, it undergoes a preprocessing pipeline to ensure high-quality input for the sentiment models essential for achieving the best results. This stage is critical for removing noise that often plagues UGC. The cleaning process involves several distinct operations:

- Artifact Removal: Specific platform-generated tags, such as "(Translated by Google)" or "(Original)", are stripped to isolate the user's actual voice.
- Normalization: The text is converted to lowercase to ensure consistency (e.g., treating "Bad" and "bad" as identical tokens) and whitespace is normalized.
- Sanitization: Non-alphanumeric characters and punctuation are removed to reduce dimensionality.
- Tokenization: The cleaned text is split into individual tokens (words) to facilitate precise analysis by the downstream models.

Finally, the processed dataset is structured into a tabular format and exported as a UTF-8 encoded Comma-Separated Values (CSV) file, serving as the clean "ground truth" for the next sentiment analysis phase.

The core logic for this cleaning pipeline and the data export is illustrated in Listing 1.

```
def clean_text(text):
    text = text.replace("(Translated by Google)", "").replace("(Original)", "")
    text = re.sub(r'\s+', ' ', text).strip().lower()
    text = re.sub(r'[^\w\s]', '', text)
    words = word_tokenize(text)
    return ' '.join(words)

# ... [Programmatic Collection Logic] ...

if raw_text and rating:
    clean_review = clean_text(raw_text)
    df.loc[len(df)] = [clean_review, rating]

```

 doi:10.20944/preprints202601.1176.v1



```
df.to_csv(output_path, index=False, encoding='utf-8-sig')
```

Listing 1: Preprocessing Pipeline and CSV

### *3.3. Phase 2: Dual-Model Sentiment Analysis Engine*

A key innovation of this methodology is the integration of two distinct model architectures to perform ABSA and recommendation extraction.

1. Supervised Model (Hugging Face BERT): We utilize the nlptown/bert-base-multilingual-uncased-sentiment model. This transformer-based model provides a strictly structured output (1-5 stars), which then mapped to "Positive" , "Neutral," or "Negative" polarity. It offers high consistency for metric comparisons.
2. Generative Model (Google Gemini 2.5 Flash): To capture nuanced feedback and generate qualitative summaries, the Gemini API integrated. Unlike the BERT model, Gemini was prompted to not only classify sentiment but also to reason and extract specific "recommendations" and "highlights" from the text.

The integration of the Generative AI client and the prompt engineering strategy is shown in Listing 2.

```
client = genai.Client(api_key=gemini_key)
for idx, row in df.iterrows():
    review_text = row['review']

    # Prompt engineering for strict classification
    prompt = f"Classify this review as Positive, Neutral, or Negative. ONLY return one word: '{review_text}'"

    try:
        response = client.models.generate_content(
            model="gemini-2.5-flash",
            contents=prompt,
            config=types.GenerateContentConfig(
                thinking_config=types.ThinkingConfig(thinking_budget=0)
            )
        )
        sentiment = response.text.strip()
    except Exception as e:
        sentiment = f"Error: {e}"
```

Listing 2: Implementation of the Gemini API call with prompt engineering to enforce structured sentiment output.

### *3.4. Phase 3: Middleware Architecture*

To facilitate seamless communication between the data processing layer and the user interface, a robust middleware using FastAPI. This API acts as a workflow engine, accepting Hypertext Transfer Protocol (HTTP) requests containing the target URL and the desired analysis model (e.g., "Gemini" or "Multilingual BERT" or both of them for all mode analysis).

Most importantly, the middleware acts as a coordinator, running independent Python modules through subprocess calls. This approach separates heavy computational workloads from the web server, keeping the system responsive. The middleware oversees the entire pipeline by:

1. Triggering the Data Collection: Initiating the programmatic extraction script.
2. Inference: Dynamically selecting the appropriate sentiment analysis script based on user input.
3. Integrating the Google Gemini API to generate a qualitative summary of the processed results.
4. Serialization: formatting the final output into JavaScript Object Notation (JSON) for the frontend.

 doi:10.20944/preprints202601.1176.v1



The implementation of this orchestration logic is shown in Listing 3.

```python
@app.post("/analyze")
async def analyze_reviews(request: AnalysisRequest):
    url = request.url
    model = request.model.lower()

    # Define paths for dataset and scripts
    dataset_name = f"google_maps_reviews_{datetime.datetime.now().strftime('%Y%m%d')}.csv"
    scraper_script = os.path.join(base_dir, "scraping_code", "google_maps_scraping.py")

    # Orchestrate the pipeline: 1. Scrape Data
    subprocess.run([python_executable, scraper_script, url, dataset_path], check=True)

    # Orchestrate the pipeline: 2. Run Selected Model
    if model == "gemini":
        script_path = os.path.join(base_dir, "sentiment_analysis_code", "gemin.py")
        subprocess.run([python_executable, script_path, dataset_path], check=True)

    # Return structured results
    return { "reviews": pd.read_csv(reviews_path).to_dict(orient="records") }
```

Listing 3: The FastAPI endpoint logic that orchestrates the data pipeline.

### *3.5. Phase 4: Visualization and User Interface*

The final phase focuses on making the analytical results accessible through a reactive web dashboard developed in Angular v20. This frontend connects directly to the middleware, consuming the JSON data to render interactive real-time visualizations.

To achieve this, the frontend application subscribes to the API's corresponding endpoint. Upon receiving the response, it dynamically transforms the raw sentiment counts into structured datasets compatible with Chart.js. This allows for an immediate visual assessment of model divergence (e.g., how many reviews Gemini marked as "Negative" versus Hugging Face).

Regardless of the selected analysis mode (Single or All-Model), the dashboard consistently renders three core visualization components to ensure immediate observability:

1. Sentiment Counts Chart: A comparative bar chart that contrasts the volume of classifications (Positive, Neutral, Negative). This allows for an immediate visual assessment of model divergence — for instance, observing how often the LLM detects "Negative" sentiment compared to the supervised classifier.
2. Star Ratings Distribution: A reference chart displaying the original numerical ratings from Google Maps. This serves as a "ground truth" baseline, enabling users to verify whether the AI's sentiment predictions align with the actual scores given by visitors.
3. Review Comparison Table: A detailed, color-coded tabular view that lists each review alongside its corresponding predictions.

The primary distinction between the two operational modes lies in the scope and generation of the Strategic Summaries:

- Single-Model Analysis: In this mode, the system focuses on a specific architecture. It displays the analysis of the reviews with a very short recommendation for the place of interest and the positives and negatives of the place corresponding strictly to the selected model. Technical Note: Even when the Hugging Face (BERT) model is selected, the system utilizes the Gemini API to generate the natural language summary, prompting the LLM to interpret and narrate the statistical sentiment data derived by the BERT model.

 doi:10.20944/preprints202601.1176.v1



- All-Model Analysis: This mode executes a parallel benchmarking. In addition to the individual summaries for both Gemini and Hugging Face, it generates a unique Combined Strategic Summary analysis. This synthesizes the specific strengths of both architectures — merging the reasoning of the LLM with the structured scoring of the SLM — to provide a small summary recommendation, "positives" and "negatives" aspects of it and the most important keywords for this site as evaluated by the Gemini model, taking into consideration both the analysis of Bert and Gemini.

The TypeScript logic handling this data transformation and chart population is detailed in Listing 4.

```
runComparison() {
  // Trigger API call for model comparison
  this.reviewsService.compareModels(this.url).subscribe({
    next: (data) => {
      this.plots = data.plots; // Receive processed metrics from FastAPI

      // Map backend data to Chart.js structure for side-by-side visualization
      if (this.plots) {
        const sentiments = ['Positive', 'Neutral', 'Negative'];

        this.sentimentChartData = {
          labels: sentiments,
          datasets: [
            {
              label: 'Gemini (Generative)',
              // Map Gemini counts to chart data points
              data: sentiments.map(s => this.plots.sentiment_counts.gemini[s] || 0),
              backgroundColor: 'rgba(54, 162, 235, 0.6)',
            },
            {
              label: 'HuggingFace (Supervised)',
              // Map HuggingFace counts to chart data points
              data: sentiments.map(s => this.plots.sentiment_counts.huggingface[s] ||
    0),
              backgroundColor: 'rgba(255, 99, 132, 0.6)',
            }
          ],
        };
      }
      this.loading = false;
    }
  });
}
```

Listing 4: TypeScript code demonstrating how backend sentiment metrics are dynamically mapped to the frontend chart components for comparative analysis.

## 4. Case Study: Archaeological Site of Mystras

To validate the efficacy of the proposed pipeline, a case study was conducted using real-world reviews from the Archaeological Site of Mystras, a United Nations Educational, Scientific and Cultural Organization (UNESCO) World Heritage site in Greece. This location was selected due to its rich cultural significance and the high volume of multi-faceted visitor feedback, which typically includes comments on historical value, physical accessibility (e.g., steep paths), and service quality. The scenario was executed in "All-Model Analysis" mode to provide a direct comparison between the Generative AI (Gemini) and the Supervised Learning (Hugging Face) models, as mentioned in section 3.

 doi:10.20944/preprints202601.1176.v1



### 4.1. Quantitative Analysis: Model Divergence

The initial output of the dashboard is the Sentiment Counts Chart (Figure 3), which visualizes the classification distribution for a sample of 10 random reviews.

The chart reveals a notable divergence in sentiment interpretation between the two architectures:

- Hugging Face (Red Bars): This model, based on the bert-base-multilingual-uncased-sentiment architecture, exhibited a strong tendency toward Positive classifications (9 reviews) and identified 1 Neutral review. It did not classify any reviews as negative in this sample batch.
- Gemini (Blue Bars): In contrast, the generative model identified 8 Positive reviews and 2 Negative reviews.

This difference points to a consistent gap in how the two models interpret sentiment when viewed at a broader level. The BERT-based model tends to smooth sentiment toward positive labels, especially when reviews carry high star ratings, whereas Gemini adopts a more cautious stance and is more willing to assign negative sentiment. These patterns at the distribution level naturally lead to a closer look at individual reviews, which is explored in the following section through a ground truth comparison.

Observing these differences is particularly valuable, as scaling the analysis to larger datasets (e.g., 100 or 1,000 reviews) would provide a clearer and more robust comparison of how different model architectures evaluate identical content; however, such an extension is constrained in this study due to API usage limits imposed by the basic edition employed.

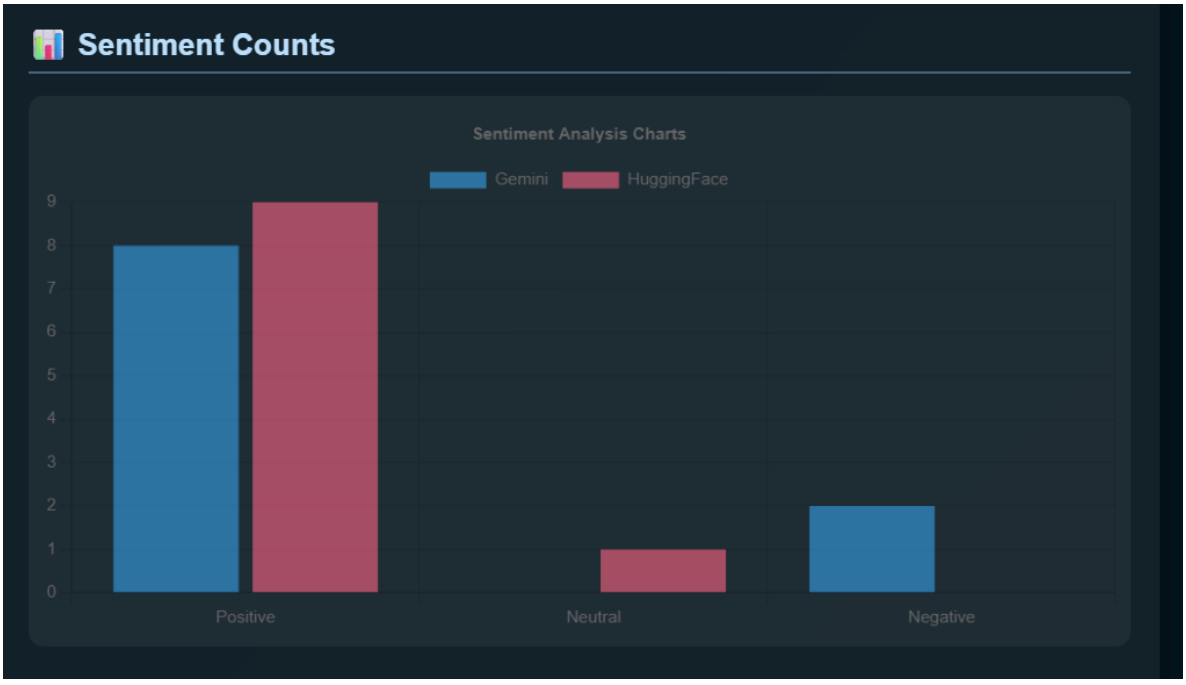


**Figure 3.** Sentiment Counts Chart comparing Gemini (Blue) and HuggingFace (Red). Note the divergence in detecting negative sentiment.

### 4.2. Ground Truth Comparison

To contextualize these predictions, the system displays the Star Ratings Distribution (Figure 4), which serves as the "ground truth" provided by the users themselves. The distribution shows a predominance of 5-star ratings (8 reviews), with single instances of 2-star and 4-star ratings.

Comparing Figures 3 and 4 reveals an interesting finding:

- The Hugging Face model's "Neutral" prediction corresponds to the 2-star review. In many sentiment datasets, 1-2 stars are negative, yet the model classified it as Neutral, potentially due to mixed language in the text ("fascinating history" vs. "making money").

 doi:10.20944/preprints202601.1176.v1



2-star review:

"It would be a nice place to see if it were taken care of better. It's like actual ruins. No signs of making it place.. lots of grass. Looks like still abandon place. The entrance fee of €20 seems too much for it. The place had its glory times and the history behind is fascinating, however, nowadays it's just a place for making money and not restoring such a nice place. Such a shame. The views from the top are just stunning. It's quite a walk uphill but it definitely worth it."

- Gemini correctly identified the 2-star review as Negative. Furthermore, it classified the 4-star review as Negative, likely picking up on specific complaints about the "closed palace" and "many stairs," demonstrating that LLMs can prioritize textual content over the numerical rating.

  4-star review: "Nice archaeological site with two monasteries well preserved. The Byzantine palace is closed, so do not go up to see it. Many stairs, a bit sloppy."

These individual examples help explain the differences observed in the overall analysis. Gemini tends to focus more on what users actually say in the text, even when that interpretation conflicts with the given star rating. In contrast, the BERT-based model appears to rely more on patterns associated with the numerical rating, which can lead to more uniformly positive (or "sanitized") predictions. This contrast displays clearly how different modeling approaches handle mixed or nuanced feedback in practice, allowing us to observe clearly the differences between the models.

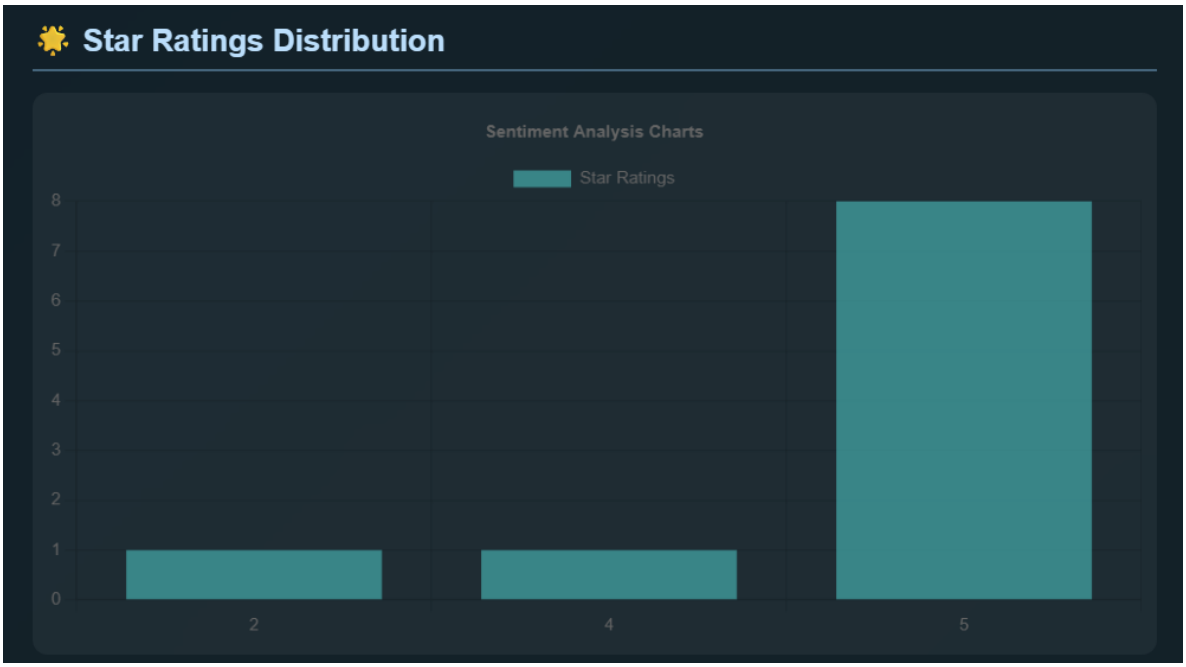


**Figure 4.** The actual Star Ratings distribution from the scraped Google Maps reviews, serving as the ground truth baseline.

*4.3. Qualitative Analysis: Automated Recommendations*

Beyond simple classification, the system successfully extracted actionable business intelligence. Figure 5 displays the Combined Summary, which synthesizes insights from both models into a coherent narrative. The pipeline successfully identified:

- Strategic Recommendations: "A must-visit archaeological site... ideal for those who enjoy active exploration".
- Specific Positive Points (Positives): The analysis identified key strengths such as "breathtaking, stunning, and glorious views," a "well-preserved archaeological site," and "helpful, welcoming, and knowledgeable staff," reinforcing the location's status as a "peak historical experience."

 doi:10.20944/preprints202601.1176.v1



- Specific Pain Points (Negatives): The system highlighted "Limited bathroom facilities at the bottom of the park" and "Byzantine palace is currently closed," providing precise feedback for site management.
- Key Highlights: It extracted specific keywords such as "Medieval Castle," "Stunning Views," and "Hiking," which are essential for marketing and tagging purposes.

This structured output demonstrates the system's ability to transform unstructured noise into clear, policy-relevant data.

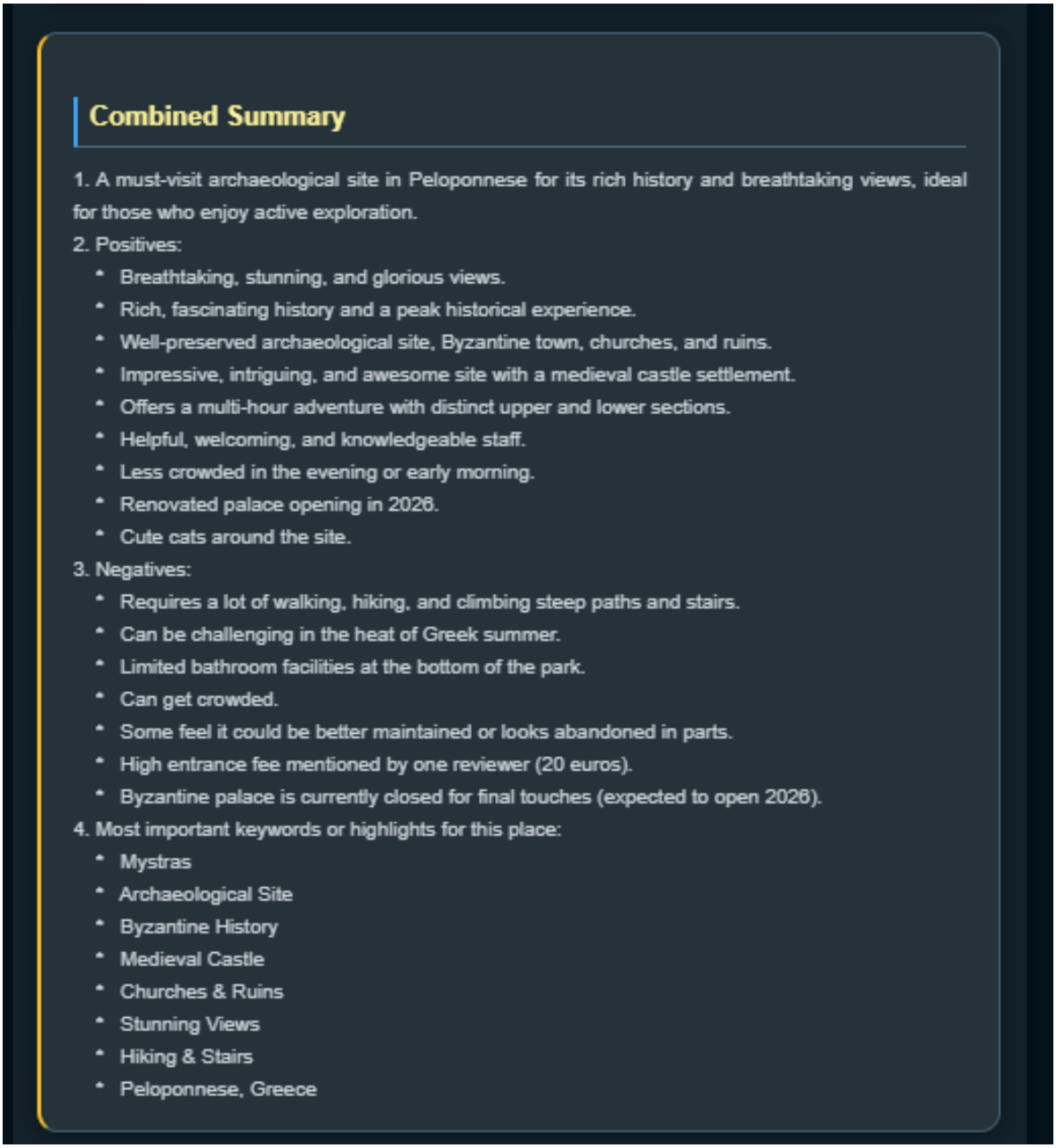


**Figure 5.** The Combined Summary generated by the system, featuring consolidated recommendations, pros/cons, and keyword highlights.

### *4.4. Granular Review Analysis*

To understand the root causes of the statistical divergence observed in Section 4.2, Review Comparison Table Figure 6 was examined. This interface displays the raw text of each review alongside its corresponding ground-truth star rating and the sentiment labels predicted by both the

 doi:10.20944/preprints202601.1176.v1



Gemini and Hugging Face models. This granular view reveals critical insights into how each model interprets "mixed" feedback:

- The "False Positive" Phenomenon (Row 1): The first review in the dataset received a 4-star rating from the user. The Hugging Face model, likely biased by the high numerical score, classified it as Positive. However, the text itself contains significant complaints: "Byzantine palace is closed... many stairs... a bit sloopy." Gemini accurately detected these linguistic cues and classified the review as Negative despite the high star rating. This demonstrates the LLM's ability to prioritize semantic content over metadata.
- The "Neutral" vs. "Negative" Distinction (Row 3): A 2-star review describing the site as an "abandon place" and criticizing the entrance fee was classified as Neutral by Hugging Face but Negative by Gemini. The supervised model likely struggled with the mixed context of historical praise ("history behind is fascinating") versus operational criticism, whereas the generative model correctly weighed the user's overall dissatisfaction.
- Consensus on Clear Positives: For unambiguous 5-star reviews (e.g., "peak historical experience," "breathtaking views"), both models achieved 100/% agreement, classifying them as Positive.

| Review | Rating | Gemini Sentiment | HuggingFace Sentiment |
|---|---|---|---|
| Nice archaeological site with two monasteries well preserved, the bizantine palace is closed so don't go up to see it. Many stairs a bit sloopy. | 4 | Negative | Positive |
| Beautiful sight highly worth visiting. As we visited in march it was not crowded at all and the entrance was for free (on a Sunday). Do note that it is quite a hike to visit the entire park. This may be challenging in the heat of the Greek summer. Bathroom facilities are limited to only at the bottom of the park/hill. | 5 | Positive | Positive |
| It would be a nice place to see if it were taken care of better. It's like actual ruins. No signs of making it place.. lots of grass. Looks like still abandon place. The entrance fee of €20 seems too much for it. The place had its glory times and the history behind is fascinating, however, nowadays it's just a place for making money and not restoring such a nice place. Such a shame. The views from the top are just stunning. It's quite a walk uphill but it definitely worth it. | 2 | Negative | Neutral |
| actually peak historical experience. the ruins are incredibly well preserved and the views are breathtaking. definitely a multi-hour adventure / hike so bring comfy / sturdy shoes. also, so many cute cats here! | 5 | Positive | Positive |
| One of the most impressive and intriguing archeological sites in the Peloponnese from the historical and christianity point of view! There are 2 gates one at the lower part and one at the upper end! You can park the car down and go up the hill and then descend on another path. You need at least 2 hours to really enjoy the place. In the evening is less crowded. The staff is wellcoming and helpful helpful! | 5 | Positive | Positive |
| I would like to give this place seven stars out of five, but I only could give five due to the technical issues. This is a sightseeing worth your while. Remarkable views, rich history, easy access, you name it. If you ever visit Peloponnesus Greece, you have to pay a visit to the Ancient Mystras! | 5 | Positive | Positive |
| What a gem. A walk through a medieval historic castle -settlement with ancient churches and glorious views of the Spartan valley. A challenging walk ,at some spots through steep stone paths. At normal pace it takes 1 h 50 minutes (including stops to admire the different churches and the small museum) to reach the mountaintop acropolis. The fully renovated palace was closed due to final touches and is opening officially April 2026. Very helpful and knowledgeable staff . Tip :You can drive up to the second entrance ,which is closer to the mountain top and then drive back to enter via the main entrance to see the museum and the larger monuments and areas. Go for it. | 5 | Positive | Positive |

**Figure 6.** The Review Comparison Table. Note the first row where Gemini detects negative sentiment in a 4-star review, while Hugging Face classifies it as Positive.

 doi:10.20944/preprints202601.1176.v1



It is important to note that in the system's "Single-Model Analysis" mode, the user would only see one of these sentiment columns. The "All-Model Analysis" presented here effectively aggregates these individual perspectives, exposing the specific linguistic triggers that cause supervised models to fail where generative models succeed.

## 5. Discussion & Future Work

The comparative analysis of the Archaeological Site of Mystras highlights a key trade-off in applying AI to cultural tourism analytics. Our findings show that supervised models such as BERT (Hugging Face) deliver fast and consistent results, but they tend to rely heavily on metadata like star ratings. As a result, they often miss subtle or "hidden" negative sentiment embedded in otherwise positive reviews. For example, a 4-star review that included complaints about accessibility was classified as Positive by the supervised model, while the Gemini LLM correctly captured the underlying frustration and labeled it as Negative.

This highlights that while LLMs excel at identifying nuanced or complex sentiments—areas where basic classifiers often fail, they are not perfect. Their performance depends heavily on how the prompt is written, and they remain prone to 'hallucinations,' where the model generates false or misleading information. In our case, even small changes in prompt phrasing could produce variations in output format, requiring a strict post-processing layer within the FastAPI middleware to maintain JSON consistency.

While Gemini produced richer and more descriptive interpretations, the Hugging Face model delivered faster, well-structured outputs that are better suited for real-time dashboard applications. In summary, a hybrid architecture provides the most robust solution for the tourism industry. By using smaller, fine-tuned models for high-speed statistical tracking and LLMs for analyzing complex qualitative feedback, stakeholders can achieve both efficiency and depth [20].

Building on this implementation, several promising directions for future research and development emerge, aimed at strengthening the system's robustness and expanding its potential societal impact. Future iterations will expand the benchmarking to include a broader range of architectures, notably more models will be added like DeepSeek and fine-tuned GPT-4 variants, to compare their performance. This approach will help assess whether open-weight models can approach the reasoning capabilities of proprietary APIs while offering potential reductions in operational costs.

At present, the analysis is static. What is proposed is an extension of the pipeline to support real-time sentiment monitoring combined with geospatial data. This would enable municipalities to visualize sentiment "hotspots" (for example, a particular museum entrance) and track temporal trends (such as sentiment dips during heatwaves), supporting more dynamic and responsive resource allocation. While this study focused on cultural heritage sites like Mystras, the methodology itself is domain-agnostic. Future work will expand the dataset to cover a variety of tourism sectors, including agrotourism and hospitality, to test how well the pipeline generalizes across different vocabularies and user expectations.

This research bridges the gap between theoretical NLP research and practical tourism management. Moving beyond static analysis, a comprehensive, end-to-end pipeline for extracting recommendations and analyzing sentiment from cultural and tourism contexts was presented. The proposed framework can be applied in real-world settings to analyze large volumes of user-generated reviews, leveraging the capabilities of modern AI and LLMs.

The contribution is defined by the following implementations:

1. A Dual-Model Framework: A comparative system that leverages both a generative LLM (Gemini 2.5 Flash) and a specialized supervised model (HuggingFace BERT Multilingual) to analyze user reviews. This facilitates an evaluation of the trade-off between the rich, reasoning-based summaries provided by LLMs and the structured, consistent scoring of SLMs. Unlike standard distinct implementations, our system features an "All-Model Analysis" mode that executes a direct, face-to-face benchmarking of both architectures against the same dataset. This design

 doi:10.20944/preprints202601.1176.v1



allows for a granular evaluation of the trade-off between the rich, context-aware reasoning provided by generative AI and the strictly structured, consistent scoring typical of fine-tuned SLMs.

2. Automated Pipeline Architecture: A robust technical architecture consisting of a Python-based scraper (utilizing Selenium for Google Maps), a FastAPI middleware for handling inference requests, and an Angular v20 dashboard for real-time visualization. This system automates the collection, cleaning and analysis of data, addressing the "data sparsity" and "preprocessing" challenges identified in previous surveys.
3. Actionable Recommendation Extraction: Beyond simple polarity, our system utilizes prompt engineering to extract specific "recommendations" and "highlights" from reviews, transforming raw text into strategic business intelligence for cultural policymakers. To demonstrate the efficacy of this approach, the methodology applied to a real-world case study: the Archaeological Site of Mystras from the google maps reviews. By analyzing visitors feedback through the pipeline, the user has the ability to compare the models (Gemini 2.5 Flash, HuggingFace BERT Multilingual) through the first mode "all model analysis" but also analyze the case of study with the model of his preference through the "single model analysis" mode.

By combining the precision of traditional NLP with the reasoning power of Generative AI, it is demonstrated that cultural institutions can move beyond simple star ratings to gain actionable insights into the specific factors shaping public perception. As digital transformation continues to reshape the tourism sector, such AI-driven tools will be indispensable for preserving heritage while modernizing the visitor economy.

**Author Contributions:** Conceptualization, N.G. ; methodology, N.G.; software, N.G.; validation, N.G.; formal analysis, N.G.; investigation, N.G.; data curation, N.G.; writing—original draft preparation, N.G.; writing—review and editing, G.T.; visualization, N.G.; supervision, G.C.; project administration, N.G., G.T.; All authors have read and agreed to the published version of the manuscript.

**Data Availability Statement:** The complete code created for this research, can be found on github: https://github.com/nikolas34g/aegean.git

**Acknowledgments:** The authors gratefully acknowledge the use of LLM tools throughout the preparation of this manuscript. These tools assisted with syntactic refinement and overall writing clarity under the direct supervision of the authors. Ultimately, all substantive content decisions, analyses, and the final composition were undertaken by the authors.

**Conflicts of Interest:** The authors declare no conflicts of interest.

 doi:10.20944/preprints202601.1176.v1



## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ABSA | Aspect-Based Sentiment Analysis |
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| BERT | Bidirectional Encoder Representations from Transformers |
| CSV | Comma-Separated Values |
| DeBERTa | Decoding-enhanced BERT with Disentangled Attention |
| GPT | Generative Pre-trained Transformer |
| HTTP | Hypertext Transfer Protocol |
| IMDb | Internet Movie Database |
| JSON | JavaScript Object Notation |
| LLMs | Large Language Models |
| NLP | Natural Language Processing |
| PaLM | Pathways Language Model |
| SA | Sentiment Analysis |
| SLMs | Small Language Models |
| UGC | User-Generated Content |
| UNESCO | United Nations Educational, Scientific and Cultural Organization |

## References


1. Alaei, A.R.; Becken, S.; Stantic, B. Sentiment analysis in tourism: Capitalizing on big data. *Journal of travel research* **2019**, *58*, 175–191.
2. Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* **2010**, *53*, 59–68.
3. García, S.; Ramírez-Gallego, S.; Luengo, J.; Benítez, J.M.; Herrera, F. Big data preprocessing: Methods and prospects. *Big data analytics* **2016**, *1*, 9.
4. Aftab, F.; Bazai, S.U.; Marjan, S.; Baloch, L.; Aslam, S.; Amphawan, A.; Neo, T.K. A comprehensive survey on sentiment analysis techniques. *International Journal of Technology* **2023**, *14*, 1288–1298.
5. Dewi, L.C.; Chandra, A.; et al. Social media web scraping using social media developers API and regex. *Procedia Computer Science* **2019**, *157*, 444–449.
6. Schouten, K.; Frasincar, F. Survey on aspect-level sentiment analysis. *IEEE transactions on knowledge and data engineering* **2015**, *28*, 813–830.
7. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* **2022**, *55*, 5731–5780.
8. Tan, K.L.; Lee, C.P.; Lim, K.M. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences* **2023**, *13*, 4550.
9. He, H.; Zhou, G.; Zhao, S. Exploring e-commerce product experience based on fusion sentiment analysis method. *Ieee Access* **2022**, *10*, 110248–110260.
10. Dragoni, M.; Donadello, I.; Cambria, E. OntoSenticNet 2: Enhancing reasoning within sentiment analysis. *IEEE Intelligent Systems* **2022**, *37*, 103–110.
11. Chifu, A.G.; Fournier, S. Sentiment difficulty in aspect-based sentiment analysis. *Mathematics* **2023**, *11*, 4647.
12. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.
13. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
14. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
15. Kalyan, K.S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal* **2024**, *6*, 100048.

 doi:10.20944/preprints202601.1176.v1



16. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.J.; Bing, L. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005* **2023**.
17. Wang, Z.; Xie, Q.; Feng, Y.; Ding, Z.; Yang, Z.; Xia, R. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339* **2023**.
18. Mughal, N.; Mujtaba, G.; Shaikh, S.; Kumar, A.; Daudpota, S.M. Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access* **2024**, *12*, 60943–60959.
19. Yue, L.; Chen, W.; Li, X.; Zuo, W.; Yin, M. A survey of sentiment analysis in social media. *Knowledge and Information Systems* **2019**, *60*, 617–663.
20. Fatemi, S.; Hu, Y. A comparative analysis of fine-tuned LLMs and few-shot learning of LLMs for financial sentiment analysis. *arXiv preprint arXiv:2312.08725* **2023**.
21. Chatzimina, M.E.; Papadaki, H.; Pontikoglou, C.; Oikonomou, N.; Tsiknakis, M. Sentiment Analysis in Greek Clinical Conversations: A Comparative Study of BERT, VADER, and Lexicon Approaches. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023, pp. 4800–4806.