Information Technology & Tourism (2025) 27:939–979
https://doi.org/10.1007/s40558-025-00334-2

ORIGINAL RESEARCH

From words to visuals: a transformer-based multi-modal framework for emotion-driven tourism analytics

Víctor Calderón-Fajardo[1,2] · Ignacio Rodríguez-Rodríguez[1] · Miguel Puig-Cabrera[3]

Received: 8 February 2025 / Revised: 6 May 2025 / Accepted: 9 July 2025 /
Published online: 22 July 2025

Abstract

Traditional tourism analytics have primarily relied on isolated sentiment analysis and image processing techniques, often failing to capture the subtle interaction between textual expressions and visual aesthetics inherent in tourist experiences. This study addresses these limitations by proposing a novel multi-modal framework that transforms textual reviews into AI-generated images using standardized prompts, thereby converting affective signals into explicit visual features. Leveraging state-of-the-art models—such as Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) for fine-grained emotion recognition and Contrastive Language–Image Pre-training (CLIP) for semantic extraction of visual attributes—our approach maps complex sentiments onto interpretable visual characteristics, integrating explainable features to uncover the underlying structure in tourist perceptions. This approach enhances classification performance and provides a transparent mechanism for understanding how distinct emotional states correspond to specific visual cues. Experimental evaluations on a dataset encompassing four diverse tourist destinations—Berlin, Dublin, Cairo, and Málaga—demonstrate high classification accuracy and robust correlations between text-derived emotions and image-based features, close to more powerful embedding methods. Significant correlations were observed between emotions and visual features, e.g., brightness and contentment, as well as between entropy and shame, indicating that our method efficiently captures the affective resonance between visual and textual modalities. Our findings underscore the transformative potential of converting textual sentiment into visual representations to facilitate more accurate, interpretable, and actionable analytics in the tourism sector. This framework suggests promising avenues for dynamic destination characterization, informed marketing strategies, and enhanced urban planning initiatives, laying the foundation for future advancements in multi-modal tourism analytics.

Extended author information available on the last page of the article

Keywords Multimodal tourism analytics · Transformer models · Text-to-Image generation · Affective sentiment analysis · Explainable AI · Destination classification

1 Introduction

The rapid proliferation of user-generated content in the tourism sector has ushered in an era where vast amounts of data—from textual reviews to user-submitted photographs—can be harnessed to gain unprecedented insights into visitor perceptions and experiences. Notwithstanding this potential, conventional sentiment analysis and topic modeling approaches have often proven inadequate in capturing the versatile and subtle emotional states embedded within these data sources (Alaei et al. 2019). While computationally efficient, traditional lexicon-based methods typically rely on predefined dictionaries and statistical heuristics that fail to account for contextual subtleties, colloquialisms, and the dynamic evolution of language (Valdez and Mehrabian 1994). Likewise, they usually require large amounts of text to produce a reliable result. Similarly, early image-based analyses frequently depended on low-level pixel statistics or basic compositional features that cannot encapsulate the semantic depth inherent in visual content (Krizhevsky et al. 2012).

Recent advances in deep learning have catalyzed the development of powerful embedding techniques, such as FastText for text (Bojanowski et al. 2017) and Inception V3 for images (Szegedy et al. 2016), which notably improve the performance of classification tasks in high-dimensional spaces, achieving good performance metrics indicating that there is an underlying data structure that allows for differentiation between classes. However, these latent representations are typically opaque, rendering them "black boxes" that provide little interpretability regarding the underlying features driving the decision-making process (Xu and Du 2019). This lack of transparency impedes the theoretical understanding of tourist behavior. It limits the practical utility of such models for stakeholders, including destination managers and policymakers, who require understanding to inform strategy.

In parallel, explainable artificial intelligence (XAI) has emerged as a promising avenue to narrow the gap between predictive accuracy and interpretability. Recent methodologies have focused on coupling deep neural architectures with interpretable layers—such as Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) for emotion recognition in text (Sanh 2019) and Contrastive Language-Image Pretraining (CLIP) for semantic estimation in images (Radford et al. 2021)—to elucidate the complex interaction between affective signals and visual attributes. Nevertheless, despite these advances, a critical methodological gap persists: effectively integrating interpretable multi-modal features remains an underexplored challenge. Existing hybrid models often struggle to balance the richness of deep feature representations with the necessity for clear, human-understandable explanations, thereby restricting their capacity to fully capture and exploit the synergies between textual and visual data.

Moreover, while individual studies have addressed aspects of sentiment analysis and visual perception separately, a dearth of research systematically correlates

emotion-laden textual features with corresponding interpretable visual attributes. For example, recent work has revealed meaningful relationships between environmental aesthetics—such as brightness and balance—and the elicitation of specific emotions like contentment or relief (Nixon et al. 2023). However, integrating these findings into a comprehensive, multi-modal classification framework that reliably differentiates among destinations remains nascent. Such an approach is imperative for advancing academic understanding and providing practitioners with robust tools capable of dynamically adjusting marketing strategies and urban planning initiatives based on real-time sentiment data.

The rationale for this study stems from the aim of demonstrating how text-to-image transformation can capture the sensations tourists describe, from the need to develop a scalable and interpretable framework that uses advanced models such as DistilBERT and CLIP to approach the performance of embedding-based methods and from the goal of providing a holistic and actionable view of tourist perceptions through robust correlations between textual emotions and image-derived cues. This perspective promises to improve the reliability of destination ranking models and offer deeper insights into the factors shaping travel experiences, leading us to pose the following research questions: RQ1: Does text-to-image transformation effectively encapsulate the emotional nuances of tourist reviews in a way that supports accurate classification and in-depth analysis? RQ2: How does a multi-modal approach that incorporates interpretable features derived from DistilBERT and CLIP compare to traditional, purely latent methods in addressing the limitations of existing frameworks? RQ3: In what ways do correlations between text-derived emotional states and specific visual attributes enrich our understanding of traveler perceptions and enhance destination characterization?

After this, Sect. 1 and 2 will detail our comprehensive methodology, including data collection, preprocessing, and feature extraction procedures for textual and visual modalities. Sect. 3 will present the experimental results, including performance metrics, classification outcomes, and correlation analyses between the interpretable features. Sect. 4 will discuss the implications of our findings, address potential limitations, and propose directions for future research in multi-modal tourism analytics. The paper concludes with Sect. 5 with some final remarks and future lines of research.

2 State of the art

2.1 Sentiment analysis in tourism: from lexicon-based methods to deep learning advances

Early studies on traveler-generated content have primarily used sentiment analysis (a Machine Learning-based technique for automatically identifying the polarity or emotion expressed in text) and topic modeling (an unsupervised data analysis method that uncovers latent themes within large corpora) to glean overarching impressions—whether positive or negative—of destinations (Alaei et al. 2019), often relying on thesauri, bag-of-words, or statistical approaches to analyze textual data. While such methods offer a coarse yet convenient measure of tourist perception, they neglect the

subtle emotional states that can significantly influence the traveler's overall experience. Automated sentiment analysis applied to tourism contexts demonstrates diminished performance compared to human raters when processing datasets characterized by increased complexity and noise (Li et al. 2022). Currently, widely used sentiment analysis modules such as Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert 2014), SentiArt (Jacobs and Kinder 2019), and similar lexicon-based tools rely heavily on pre-established dictionaries or thesauri that map specific words to certain polarity scores. While these methods can effectively capture general sentiment trends, they fail to adapt to the continuous evolution of language or require long texts to be efficient (Hartmann et al. 2022). Consequently, they may misinterpret slang, neologisms, or colloquial expressions with context-dependent meanings. Moreover, specific registers or domains, such as specialized tourist jargon, can exhibit vocabulary gaps not covered by conventional lexical resources (Pryma 2023). As a result, these limitations can yield incomplete or misleading sentiment evaluations in real-world applications. In addition, these methods are often helpful with long texts, capturing sentiment across entire books or long chapters, failing with short texts such as opinion reviews of a few lines.

More advanced deep learning methods, including word embeddings such as Word-2Vec and Global Vectors for Word Representation (GloVe) (which represent words as dense numeric vectors capturing semantic relationships) or transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT) (Rezaeinia et al. 2019), (which employ self-attention mechanisms to analyze words in context) have surfaced recently as powerful tools for capturing richer semantic information, thus driving improved classification performance across multiple tasks in tourism analytics (Meng et al. 2024). However, some approaches often present an interpretability dilemma, with their latent feature spaces acting as "black boxes" (Xu and Du 2019), (meaning it is challenging to explain how the model derives its predictions) making it difficult for stakeholders—destination managers or policy-makers—to comprehend precisely which textual cues influence model predictions, while other approaches in text analysis have proven to be useful for semantic estimation of concepts. Another complication is that many neural language models struggle with domain adaptation unless fine-tuned on travel-specific corpora, which may not be readily available or large enough to ensure reliable generalization (Singhal et al. 2023). Moreover, while some researchers have explored hybrid techniques that incorporate external knowledge bases or domain ontologies (structured representations of concepts and relationships), such methods can be challenging to maintain and update, especially if new travel-related phrases or colloquialisms emerge rapidly (Noira et al. 2021).

2.2 Image-based approaches and generative models for multi-modal tourism research

Parallel investigations into image-based studies have aimed to capture tourists' emotional perceptions by analyzing user-generated photographs from platforms such as Instagram or Flickr. Through convolutional neural networks (specialized at detecting patterns in image data) for scene recognition or visual sentiment extraction (Cilkin

and Çizel 2022), these efforts have linked specific motifs (e.g., beaches, nightlife) to expressed feelings. However, a fundamental issue remains: genuine photographs may not always mirror travelers' subtle emotions in written reviews, especially when language barriers and cultural subtleties come into play (Marder et al. 2019). AI-generated imagery offers a promising alternative to address this gap by transforming textual prompts derived from travelers' descriptions into standardized, language-agnostic visuals (Fu et al. 2024). This approach preserves the emotional essence behind each review and transcends potential linguistic or contextual disparities, ensuring a consistent visual representation of subjective experience. Furthermore, the integration of generative imagery alleviates concerns about insufficient user-uploaded content for rarer or emerging destinations that may not yet have a substantial photographic footprint. By using textual descriptions rather than real-world photos, studies can also systematically explore hypothetical scenarios—such as projecting how a yet-to-be-developed resort might evoke certain emotions—without relying on existing image repositories (Wang et al. 2020).

Advances in generative language-and-vision models (AI systems that create images or text by learning from massive paired datasets) have further amplified these opportunities, as systems like DALL-E interpret textual input and produce diverse images that approximate the essence of the provided descriptions. Meanwhile, multimodal embeddings (representations that combine textual and visual information into a single feature space) align textual and visual tokens in a joint space, enabling a range of tasks from zero-shot classification to cross-modal retrieval (Wu et al. 2023). For instance, a tourist's description mentioning "intense sunlight and bustling night markets" might yield an image reflecting those exact attributes, providing researchers with a consistent visual proxy. This capacity for automated generation helps circumvent the frequent unevenness found in user-uploaded content, expanding the scope of tourism analytics to more systematically capture how travelers might perceive or imagine different destinations. Notably, such generative models can also foster inclusivity by offering multiple stylistic variations of an image—allowing individuals from various cultural backgrounds to relate more directly to distinct visual representations of the same underlying textual concept (Zaino et al. 2022).

In parallel, emotion recognition in the text has made substantial strides beyond traditional sentiment polarity. Models like DistilBERT (a compact variant of BERT that still leverages deep contextual understanding) now handle multi-label settings that capture a richer array of emotions, which can be pivotal in shaping a traveler's destination choices (Sanh 2019). However, combining this emotional granularity with equally interpretable visual features has remained a relatively unexplored area. While XAI initiatives (explainable AI methods aimed at clarifying the decision-making process of complex models) have sparked interest in revealing which inputs drive neural network decisions (Saeed and Omlin 2021), a significant portion of tourism research relies on less interpretable embedding techniques (vector representations that capture semantic or structural properties of data), such as FastText for text or Inception V3 for images. These methods can yield strong classification results in tasks like destination differentiation, indicating a discriminating structure in the data and marking a maximum in the classification performance. Still, the latent dimensions (hidden features that emerge within the model but remain difficult for humans to interpret)

remain opaque, limiting direct insights into how or why a destination evokes specific responses. Additionally, even when emotion detection is performed accurately at a granular level, researchers face the challenge of mapping these detected emotions to specific physical or cultural elements of a destination if only real-world images are used—reinforcing the value of synthetic visuals that can abstractly represent key motifs (Folino et al. 2024).

2.3 Gaps in multi-modal tourism research and proposed direction

Despite these methodological advances, several gaps persist within multi-modal tourism research. Many prior studies predominantly incorporate observational data, focusing on photographs voluntarily uploaded by travelers, thereby risking selection bias in the depicted content (Marder et al. 2019). On the other hand, although occasional correlation studies examine alignments between textual sentiment and the imagery travelers post (Zhao et al. 2019), a more comprehensive framework linking emotion-laden textual features and interpretable visual attributes is seldom seen (Al-Tameemi et al. 2023; Wen and Xu 2024). Moreover, the question of whether such interpretable features, when fused, facilitate improved classification of tourist destinations—thus enhancing data-driven characterization (an AI-based process that uses data patterns to describe or classify a target variable)—remains underexplored. From a methodological perspective, challenges arise when attempting to consolidate quantitative metrics from disparate data modalities since text and image features can differ in dimensionality and the nature of the information conveyed. Researchers thus require robust data-fusion strategies (integrating multiple sources of data) or attention-based architectures (deep learning models that focus selectively on the most relevant parts of the input) that can meaningfully merge textual emotions and visual cues into a single decision-making pipeline (Yin et al. 2023). As tourism research increasingly uses cross-modal and explainable methods, future studies could use new generative approaches to create immediate, emotion-based depictions of possible or changing travel situations. This will further advance destination analytics. (Regan et al. 2024).

3 Methodology

This methodology begins by gathering a balanced set of user-generated tourist reviews for Berlin, Dublin, Cairo, and Malaga, ensuring anonymity through careful preprocessing of location-specific references. Each textual review is transformed into a synthetic image with a standardized prompt and DALL-E, yielding a complementary visual dataset. Next, both text and images undergo feature extraction using embedding methods and, in parallel, interpretable sets of attributes. Some classic Machine Learning (ML) classifiers are trained and evaluated using five-fold cross-validation, assessing performance with Area Under the Curve (AUC), Accuracy, F1-score, and confusion matrices in all combinations of the four problems posed (text or images, both with embedding features or explainable features). Permutation Feature Importance is subsequently applied to rank the most decisive features for city

classification. At the same time, correlations between the top emotion-focused and visually focused attributes reveal how specific affective states align with particular aesthetic or compositional cues. Finally, combining the five most salient textual and five most salient visual features further refines classification performance, highlighting the synergy between emotional and visual signals.

An outline of the methodological steps is shown in Fig. 1.

All computations are performed in Python using specialized data science libraries.

3.1 Data collection: text data

The corpus for this study comprises 180 curated tourist reviews, evenly distributed across Berlin, Dublin, Cairo and Málaga, four destinations selected to maximise cultural, climatic and experiential contrast. Berlin contributes the voice of a post-reunification capital whose museums, club scene and street art epitomise contemporary Central Europe. Dublin adds perspectives informed by Gaelic storytelling traditions and a lively pub culture that help define Western European hospitality (James and James 2020). Cairo, with pyramids, bazaars and a population exceeding twenty million, offers insights from a Middle-Eastern and North-African metropolis with continuous urban life since antiquity (Rico 2021). Málaga supplies a Mediterranean vantage point rooted in Andalusian heritage, vibrant coastal leisure and a mature tourism infrastructure that attracts year-round sun-seekers (Ruiz et al. 2019). Together, these four cities create a matrix broad enough to test whether traveller sentiment differs by historical depth, language family, latitude or dominant trip purpose, yet compact enough to allow close reading of individual narratives.

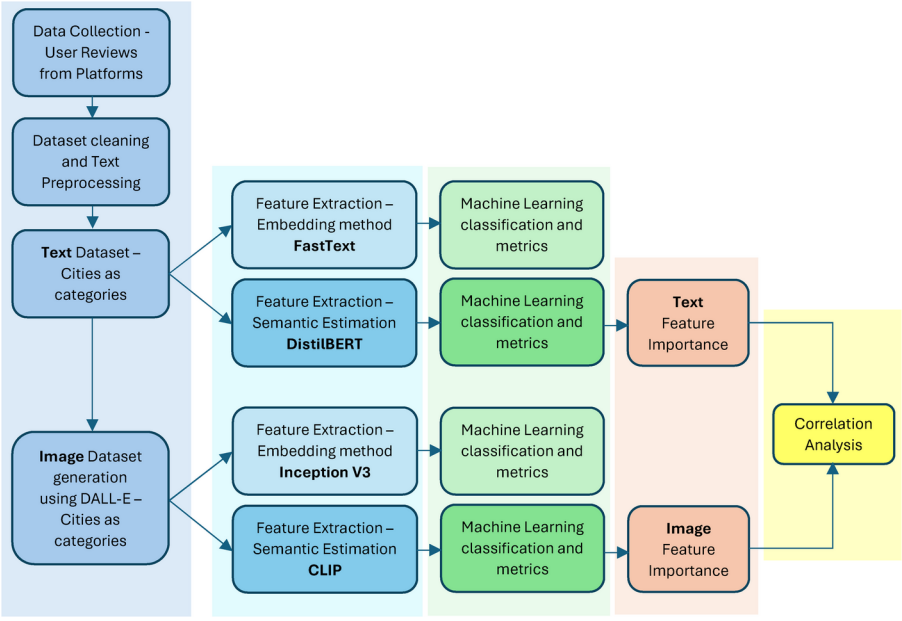

Fig. 1 Methodological outline

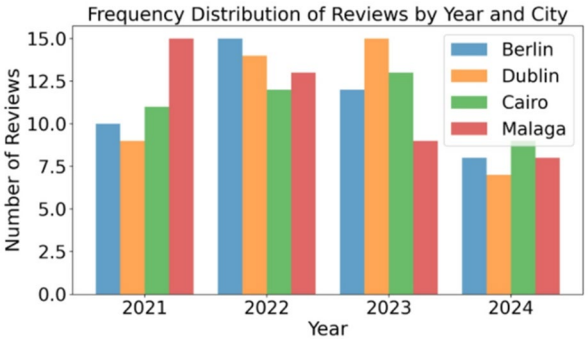

Fig. 2 Frequency distributions of reviews over the years

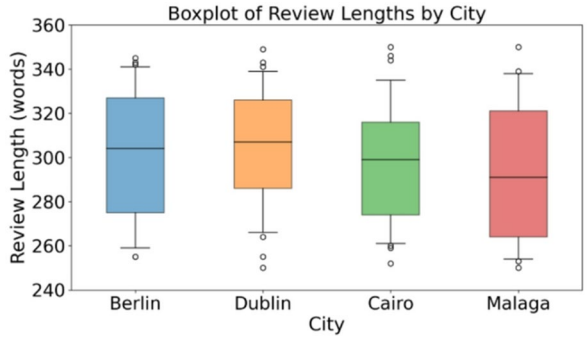

Fig. 3 Boxplot distribution of word numbers per review of cities

Temporal scope was deliberately limited to comments posted from 2021 onward to minimise distortions tied to early-pandemic uncertainty; Fig. 2 shows the resulting histogram, which clusters review activity between 2021 and 2024, a period in which border restrictions had largely stabilised yet travellers' memories remained fresh. Although most reviews were authored in English, a handful in German, Spanish or Arabic were translated when they offered especially vivid or emotionally nuanced descriptions. Reviews were harvested from four high-traffic forums—TripAdvisor, Expedia, Booking.com and Lonely Planet—and from "traveller story" repositories hosted by "VisitBerlin", "Fáilte Ireland", the Egyptian Tourism Authority and "Turismo y Planificación Costa del Sol" (Oliveira et al. 2020). Priority was given to sources with well-established reputations for reliability and high user engagement, thereby maximizing the representativeness of tourist experiences (Rasul et al. 2024). The raw crawl yielded 306 unique texts.

A several-layer quality-control pipeline refined the dataset. First, potential duplicates were detected 27 cross-posted or near-identical comments were eliminated, preventing overrepresentation of copy-and-paste hotel praise. Language was verified using fastText; 14 texts whose probability of being English fell below 0.90, or that displayed machine-translation artefacts, were removed. Next, length normalisation enforced a corridor of 250 to 350 words: 19 overly brief pieces failed to supply sufficient narrative context, while 9 lengthy digressions risked topic drift. The surviving reviews cluster tightly around a median of about 300 words, as the boxplot in Fig. 3 confirms, providing comparable document sizes for downstream embeddings.

Subsequently, topical purity was assessed by semantic similarity and projecting each review's Term Frequency–Inverse Document Frequency (TF–IDF) vector encompassing attractions, sensory adjectives, service descriptors and activity verbs; cosine similarity below 0.20 indicated material dominated by logistics such as visa issues, flight cancellations or baggage delays, leading to the exclusion of 9 items. Sentiment skew was mitigated by computing VADER compound polarity for every text and standardising it within each city; 23 comments lying beyond 3 standard deviations, or whose superlatives exceeded 5 per cent of tokens, or that presented hype or rage statements such as "Best holiday ever!" or "Worst place on earth" were discarded. These extreme opinions, while genuine, tend to distort aggregate sentiment and are often associated with covert marketing or cathartic venting rather than balanced reflection.

Finally, commercial spam and toxicity were screened. Regular-expression scans flagged comments containing two or more external URLs, three or more repeated brand references or overt contact information, criteria aligned with TripAdvisor's March 2024 anti-spam update and the European Commission's guidelines on user-generated content; 14 posts met at least one condition and were removed. Offensive language detection relied on Hatebase v4 and the Canadian Institute for Cybersecurity of the University of New Brunswick (UNB-CIC) abuse taxonomy; eleven texts containing insults, hate slurs or explicit profanity were confirmed by dual annotators and excluded to protect analytic integrity and comply with ethical standards (Wattanacharoensil and La-Ornual 2019). In total, 126 reviews—roughly 41% of the initial pool—were filtered out, leaving 180 high-quality narratives.

To prevent inadvertent labeling of the city being reviewed, all proprietary nouns indicative of location—such as district names, iconic monuments, or other geographic markers—were systematically removed or replaced with neutral placeholders during text preprocessing. Because user profiles commonly rely on partial or pseudonymous information, anonymity was inherently preserved, and we captured only the textual portion of each review without storing any identifying data. The anonymization procedure is essential to mitigate potential confounding variables in subsequent text classification and embedding analyses. A balanced corpus of tourist reviews was constructed through meticulous curation and comprehensive data cleaning, encompassing a representative spectrum of experiences and sentiments for each of the four selected cities.

3.2 Image data generation

In this work, image generation is achieved using a prompt to the DALL-E generative language and vision model through the Generative Pre-trained Transformer 4 (GPT-4) interface (Nichol et al. 2021). Table 1 below illustrates the standardized prompt format. Each prompt begins with a directive to generate "a single, unified image that accurately represents the sentiment and key elements of the review," followed by instructions regarding forbidden text or explicit naming. This approach encourages the system to capture the symbolic and emotional core of the tourist's experience without directly revealing the geographic context (Hao et al. 2022).

From a technical standpoint, DALL-E operates on a neural architecture that encodes the input text into a latent representation and then decodes this representa-

Table 1 Standardized prompt for unbiased image generation

Prompt
Please generate a single, unified image that accurately represents the sentiment and key elements of the following user review about visiting "the city." Do not include any text, names, letters, or labels within the image. Replace specific references to landmarks or locations with neutral descriptors (e.g., "large tower" or "wide river"). Summarize the entire impression of the comment in one cohesive visual that captures the traveler's overall experience and atmosphere without revealing the city's identity. Avoid any hidden bias or culturally sensitive misrepresentation. The final result must be a single image, not a collection of multiple images, representing the user's perspective as conveyed in the review.

tion into an image based on learned patterns from vast amounts of text-image pairs (Nichol et al. 2021). GPT-4, primarily known for its language processing prowess, interfaces with the DALL-E model to interpret the user's prompt and structure it into a suitable format for the image generation request. This synergy leverages large-scale transformer-based networks, wherein multi-head self-attention mechanisms interpret and weigh different aspects of the input text, mapping them to visual features. Once the text is embedded into latent space, the decoder iteratively refines the image through a learned sequence of operations guided by probabilities of pixel (or token) generation at each step (Wang 2024).

In order to mitigate the potential for cultural bias arising from the training corpus on which DALL·E relies, we note that its broad and sometimes contrasting data sources can function as a "melting pot," reducing the likelihood of overt prejudices. Nonetheless, a manual review of the generated images was performed to confirm their alignment with the study objectives and to avoid inadvertently perpetuating harmful stereotypes.

3.3 Feature extraction by embeddings methods

Embedding-based techniques can generate large sets of features that, while unexplainable, often yield high classification accuracy from text and images. This dual-modal framework bypasses explainable feature extraction methods, mitigating biases and reducing development time (Vargas et al. 2024). Accordingly, we first measure these embeddings' raw predictive power for city identification across user reviews and AI-generated images to establish an indicative upper bound on ranking metrics and subsequently see how close we can get to this limit with XAI.

3.3.1 Feature extraction for text

Textual embeddings were obtained from FastText, a popular open-source library developed by Facebook's AI Research team (Bojanowski et al. 2017). Unlike other word embedding models (e.g., Word2Vec), FastText treats each token as a composition of character n-grams. This architectural choice allows the model to better handle rare words and out-of-vocabulary tokens by breaking them into subword units. FastText embeddings are thus particularly robust for handling informal or domain-specific texts that could include misspellings, slang, and language variations. Unlike earlier embedding models such as Word2Vec (Mikolov et al. 2013) or GloVe (Pennington et al. 2014), We also considered newer sentence-level embeddings, including

Sentence-BERT (Reimers and Gurevych 2019) and the Universal Sentence Encoder (Cer et al. 2018). However, we ultimately opted for FastText due to its well-established efficiency, comparatively straightforward training requirements, and proven capacity to serve as a reliable baseline in various NLP tasks.

The preprocessing pipeline begins by applying tokenization, whereby each review is broken down into tokens (words or subword units) by removing punctuation and segmenting whitespace. This is followed by stopword removal, in which common English words like "the," "and," or "or" are filtered out to reduce noise and enhance the relevance of the remaining tokens. Once the text is preprocessed, feature vectors are generated using FastText embeddings, where each token in a review is represented by a fixed-dimension vector (often 300 dimensions). An aggregation step is applied to derive a single embedding per review—most commonly by taking the mean of all token embeddings, although weighted schemes may also be employed. This procedure produces a single high-dimensional vector for each document, capturing the semantic content of the text in a compact form (Bojanowski et al. 2017).

3.3.2 Feature extraction for images

Visual data from AI-generated images were encoded using Inception V3 (Xia et al., 2017), a convolutional neural network architecture developed by Google (Szegedy et al. 2016). Inception V3 employs a series of inception modules, factorized convolutions, and auxiliary classifiers to capture hierarchical visual features efficiently. Training on a large-scale dataset—ImageNet—endows the model with a broad "universal" feature extractor that can generalize to various image classification tasks. Inception V3 (Szegedy et al. 2016) was chosen for image feature extraction over other widely recognized architectures such as Visual Geometry Group (VGG) (Simonyan and Zisserman 2014), ResNet (He et al. 2016), or EfficientNet (Tan and Le 2019) due to its consistent performance, relatively modest computational requirements, and established status as a standard baseline in computer vision. Although more recent models like EfficientNet may reach higher accuracy on specific benchmarks, the extensive documentation and straightforward implementation of Inception V3 make it an ideal reference point for this study's baseline performance.

Each AI-generated image undergoes a standardized preprocessing phase to ensure compatibility with the Inception V3 architecture. First, the image is resized to 299×299 pixels, the default input shape required by the model. Subsequently, pixel intensity values are either rescaled to the [0,1] range or standardized according to the statistical distribution of the original training set. Following preprocessing, we perform embedding extraction by tapping into the penultimate layer of Inception V3 rather than the network's final classification layer. The penultimate layer's output, typically a 2048-dimensional vector, then serves as the final image embedding, succinctly reflecting the most salient visual attributes learned during training.

3.4 Feature extraction by explanatory methods

3.4.1 Feature extraction for text with DistilBERT for emotion recognition

DistilBERT is a distilled version of the BERT model (Sanh 2019), designed to achieve comparable performance with a significantly reduced computational footprint. This is accomplished through knowledge distillation, where a smaller "student" model learns to replicate the behavior of a larger, pre-trained "teacher" model by minimizing a Kullback-Leibler divergence loss over the teacher's output probabilities (Wu et al. 2024). Despite its reduced size, DistilBERT preserves many of BERT's core linguistic and contextual capabilities (Devlin et al. 2019), making it a suitable foundation for emotion recognition tasks that demand subtle semantic understanding (Sanh 2019). Although full-scale models such as GPT (Radford et al. 2019) or BERT can offer deeper contextual understanding, they typically demand substantial fine-tuning resources. DistilBERT provides a more balanced approach, capturing nuanced emotional or contextual signals in real time or under constrained conditions. Alternative models like Robustly Optimized BERT Approach (RoBERTa) (Warstadtet al. 2020) may demonstrate comparable or superior performance metrics, but the distilled architecture selected here strikes a practical trade-off between speed, accuracy, and interpretability.

3.4.1.1 Text preprocessing Before the texts are fed into DistilBERT, a rigorous preprocessing pipeline ensures they align with the model's tokenization scheme and input constraints. This pipeline typically includes normalization (e.g., lowercasing, if applicable to the model's case sensitivity), removing extraneous symbols or markup, and splitting into subword tokens via WordPiece (Song et al. 2020). These subword tokens help handle unseen terms and morphological variations, a critical feature for domains like social media or user reviews where non-standard word usage is common.

3.4.1.2 Feature generation via DistilBERT Once preprocessed text is tokenized, DistilBERT encodes the sequence of tokens into high-dimensional contextual embeddings. Through a series of transformer layers, each token representation is enhanced with a global context, allowing the model to obtain semantic traces, long-range dependencies, and details in an emotional tone. For emotion recognition, the final hidden states are typically fed into a classification head, often comprising one

or more fully connected layers that map the distilled embeddings to a probability distribution over emotion categories.

3.4.1.3 Emotion-specific output dimensions In this study, DistilBERT is trained on GoEmotions, an emotion recognition dataset (Demszky et al. 2020) that includes the following set of 28 potential emotional categories:

1. Sadness: A sense of sorrow, grief, or emotional pain.
2. Joy: Happiness, pleasure, or positive exhilaration.
3. Love: Affection, deep fondness, or attachment toward others or concepts.
4. Anger: Irritation, hostility, or resentment triggered by perceived grievances.
5. Fear: Apprehension, dread, or concern about potential threats.
6. Surprise: Sudden astonishment or unexpected wonder in response to new events.
7. Trust: Reliance, confidence, or belief in someone or something.
8. Anticipation: Expectation, eagerness, or readiness for future events.
9. Disgust: Revulsion, repulsion, or moral aversion to certain stimuli.
10. Guilt: Signals remorse or responsibility for wrongdoing or mistakes.
11. Confusion: Uncertainty, perplexity, or mental disorientation.
12. Gratitude: Thankfulness, appreciation, or recognition of kindness.
13. Pride: Satisfaction in achievements or self-worth, often socially oriented.
14. Envy: Jealousy or covetousness regarding others' advantages or possessions.
15. Optimism: Hopefulness, positivity, or confidence about future outcomes.
16. Boredom: A lack of interest, monotony, or tediousness in one's environment.
17. Shame: A sense of humiliation or distress over perceived personal failings.
18. Loneliness: Feelings of isolation, disconnection, or lack of companionship.
19. Hope: Forward-looking expectation of positive change or improvement.
20. Relief: Easing anxiety, worry, or tension once adversity subsides.
21. Disappointment: When expectations fall short, engendering feelings of letdown or regret.
22. Curiosity: A desire for knowledge, exploration, or novel experiences.
23. Excitement: Eagerness, heightened interest, or anticipation of a stimulating event.
24. Nostalgia: Sentimental longing or wistfulness for past experiences or contexts.
25. Contentment: Satisfaction, comfort, or ease in present circumstances.
26. Resentment: Lingering bitterness or indignation due to perceived unfairness or harm.
27. Frustration: Blocked goals, obstacles, or repeated disappointments.
28. Admiration: Respect, esteem, or veneration for qualities, achievements, or attributes.

By framing emotion detection as a multi-label classification task, the model can attribute varying degrees of likelihood to each emotion. For instance, a single text passage may concurrently express gratitude and relief or anger and frustration, requiring the model to quantify the probability of each emotion independently.

3.4.1.4 Semantic estimation and explanation DistilBERT is designed to enable the capture of complex semantic and emotional patterns that might be overlooked by simpler embedding-based methods (Ng et al. 2023). The output probabilities for each emotion class effectively serve as interpretable features, providing insight into which affective dimensions are most relevant in a given review. This property facilitates downstream analyses, enabling researchers to track how certain emotional states correlate with specific cities, experiences, or even aspects of the AI-generated images.

3.4.2 Visual feature extraction for images using CLIP

CLIP is a multi-modal neural network architecture developed by OpenAI, designed to align textual and visual representations within a shared latent space (Radford et al. 2021). While its most prominent use cases involve matching captions to images or zero-shot classification, CLIP's encoder modules can also be leveraged to generate rich embeddings from images alone. These embeddings capture low-level visual patterns (e.g., color gradients and edges) and high-level semantic cues (e.g., object types and contextual relationships), making them suitable for sophisticated feature extraction and downstream analyses. CLIP (Radford et al. 2021) was incorporated to project textual and visual signals into a shared embedding space, simplifying cross-modal comparisons in ways not as readily achieved with networks like Residual Network (ResNet) (He et al. 2016) or EfficientNet (Tan and Le 2019). While alternative multi-modal frameworks, including Vision-and-Language BERT (ViLBERT) (Lu et al. 2019) or VisualBERT (Li et al. 2019), also aim to merge text and image representations, CLIP's training objective and extensive pretraining dataset have proven especially advantageous for capturing both modalities within a cohesive, interlinked vector space. This cross-domain alignment is central to linking text-derived emotions with image-based features, a key objective in our research.

3.4.2.1 Image preprocessing Before extracting embeddings through CLIP, each DALL-E-generated image undergoes a standardized workflow to ensure compatibility with the model's input specifications. The main steps include resizing images to meet the resolution constraints set by CLIP's vision transformer and normalizing pixel intensities to match the training distribution used by the pre-trained CLIP model (Radford et al. 2021). This ensures that the subsequent embeddings accurately reflect the image's intrinsic features rather than artifacts introduced by preprocessing inconsistencies.

3.4.2.2 Feature generation via CLIP Unlike purely latent embeddings, the features enumerated below correspond to interpretable visual properties. Each dimension is designed to capture an attribute or aesthetic quality evident within the AI-generated image. The resulting measurements allow researchers to differentiate images holistically—bright versus dark or colorful versus monochrome—and to analyze more subtle compositional elements, such as symmetry. This interpretability parallels the

emotion categories in DistilBERT-based text analysis, thus providing a symmetrical structure for analyzing and comparing text and image data.

3.4.2.3 Visual-Specific output dimensions Once preprocessing is complete, each image is passed into CLIP's vision encoder, which typically employs a transformer-based architecture. The encoder outputs a high-dimensional embedding capturing visual semantics learned from extensive training on image-text pairs. This global embedding can be refined or dissected to yield more interpretable feature sets, such as the following 27 visual characteristics identified for this study:

1. Brightness: The average luminance or perceived lightness across the image.
2. Contrast: The difference in luminance or color allows distinct image elements to stand out.
3. Exposure: The extent to which the image appears over- or under-exposed, reflecting incident light levels.
4. Dynamic_range: The ratio between the darkest and brightest parts, indicating the breadth of visible tones.
5. Saturation: The intensity or purity of the image's colors.
6. Hue_mean: The average dominant hue or color value calculated over the entire image.
7. Hue_dominant: The single color hue that occupies most of the image.
8. Colorfulness: A measure of the overall variety and vibrancy of colors present.
9. Monochrome: The degree to which the image is grayscale or lacking in color diversity.
10. Sharpness: The clarity of edges and fine details reflecting how in-focus the image is.
11. Blur: The presence of smooth, unfocused regions due to motion, depth of field, or processing effects.
12. Texture_complexity: The intricacy or granularity of surface patterns.
13. Edge_density: The concentration of detected edges or boundaries throughout the image.
14. Gradient_magnitude: The intensity of transitions between color or luminance values.
15. Entropy: A statistical measure of the complexity or unpredictability in pixel distribution.
16. Symmetry: The degree of reflective or rotational balance in the arrangement of visual elements.
17. Rule_of_thirds: The extent to which key subjects align with compositional grid lines that divide the frame.
18. Center_of_mass: The weighted average location of all visual components, indicating overall layout balance.
19. Aspect_ratio: The proportional relationship between the image's width and height.

20. Balance: The level of color temperature and hue adjustment to make the scene's colors look natural.
21. Shadow_density: The intensity and spread of shadowed regions.
22. Highlight_density: The prominence and extent of bright, potentially overexposed areas.
23. Light_directionality: The primary orientation of illumination, inferred from shadows and highlights.

By inferring these features, we obtain a structured representation beyond a simple CLIP embedding vector and pinpoint specific image characteristics relevant to aesthetic and compositional analysis.

3.4.2.4 Semantic estimation and explanation CLIP's specialized ability to learn visual-semantic alignments can also be leveraged for higher-level semantic estimation (Gandelsman et al. 2023), bringing explanatory power. In combination with the extracted numerical features (e.g., colorfulness, sharpness), CLIP-based semantic estimation enables a dual-level interpretation—where global, latent semantics complement explicit, human-interpretable attributes. As a result, this layered approach to feature extraction and semantic matching offers a powerful toolkit for examining how generated images reflect the sentiments, themes, or contexts outlined in corresponding textual reviews.

3.5 Classifications, performances, and features ranks

3.5.1 Machine learning classifiers

Once we have generated vector embeddings for both text (FastText) and images (Inception V3), as well as explainable features from text (DistilBERT) and images (CLIP), the next step involves feeding these four different scenarios—separately—into a dedicated classification module. In this study, we explore and compare three principal classifiers: Support Vector Machine (SVM) (Cortes and Vapnik 1995), Logistic Regression (LR) (Lever et al. 2016), and a Neural Network (NN) Multi-Layer Perceptron. SVM is adept at handling smaller datasets with high-dimensional feature representations, maintaining robustness against the "curse of dimensionality" even under limited sample sizes (Köppen 2000). By contrast, LR provides an interpretable baseline with fewer trainable parameters, though it may struggle with highly non-linear decision boundaries. Meanwhile, an NN-MLP can capture complex patterns through multiple hidden layers but generally requires larger training sets. On the contrary, although tree-based methods (e.g., Random Forests or Gradient Boosting) often demonstrate strong performance, they commonly demand substantial hyperparameter tuning to manage correlated or continuous embeddings and can lose subtle information through iterative splits (Rabinowicz and Rosset 2022) and therefore have not been considered in this work. By comparing SVM, LR, and NN MLP under consistent experimental conditions, we ensure that any observed variations in

classification metrics are predominantly attributable to the distinct feature representations—rather than inherent differences in the learning algorithms.

Regarding SVM, we adopt a Radial Basis Function (RBF) kernel to accommodate non-linear decision boundaries, which is particularly valuable for high-dimensional embeddings. The principal hyperparameters include C, which balances margin maximization against misclassification costs, and γ (gamma), governing the kernel's width. We allow up to 500 iterations; at this point, the solver terminates if convergence is not yet achieved. LR serves as a robust, interpretable baseline when the dataset is small. Its linear decision boundary and fewer trainable parameters reduce the risk of overfitting while still providing solid performance in moderately high-dimensional spaces. We apply the scikit-learn implementation with an L2 penalty, tuning the regularization parameter C to balance complexity against misclassification. A maximum of 500 iterations is similarly used to ensure adequate model fitting. We include an MLP that can capture complex, non-linear relationships for completeness. We configure a hidden layer (150 neurons) with ReLU activation for efficient gradient propagation and a softmax output layer of four neurons—one for each city class. Using SGD as the solver, we set an alpha (L2 penalty) to mitigate overfitting and a maximum of 500 iterations for convergence. While MLPs can yield strong results, they typically require larger datasets to avoid overfitting, making them a valuable comparison point alongside SVM and LR.

A stratified five-fold cross-validation (CV) approach is used to evaluate the model under rigorous conditions. The dataset is split into five roughly equal partitions: each partition successively serves as the validation set, while the other four are used for model training. This method minimizes overfitting and provides robust performance estimates by iterating through all possible validation folds.

3.5.2 Performance metrics

To evaluate model performance, we employ primary metrics—Area Under the ROC Curve (AUC), Accuracy, F1-score, and confusion matrices—each providing a different perspective on classification efficacy (Grandini et al. 2020).

3.5.2.1 Area under the ROC curve (AUC) AUC indicates how well the model ranks positives versus negatives across various classification thresholds. An AUC of 1.0 reflects perfect separability, whereas an AUC of 0.5 denotes random guessing. Since we have four classes (one for each city), we often compute the macro-averaged AUC by aggregating one-vs.-rest curves for each city and averaging the result (Grandini et al. 2020).

Accuracy:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

This metric measures the proportion of samples for which the predicted city matches the actual label (Eq. 1). While straightforward, accuracy may be misleading if a class imbalance exists (Grandini et al. 2020).

F1-score:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2}$$

The F1-score harmonizes precision (the proportion of predicted positives that are positive) and recall (the proportion of actual positives correctly identified) (Eq. 2), offering robustness in scenarios with disproportionate class distributions. A higher F1 implies a balanced class performance (Grandini et al. 2020).

3.5.2.2 Confusion matrices A confusion matrix provides a comprehensive view of classification performance by mapping predicted labels against the actual labels in a square matrix. This four-class problem is a 4×4 grid where each row represents the actual city, and each column corresponds to the predicted city. Diagonal entries—where predicted and actual labels coincide—indicate correct classifications, while off-diagonal cells reveal misclassifications, such as instances from one city erroneously attributed to another. By inspecting the distribution of errors and successes across all classes, researchers can identify patterns of systematic mislabeling, assess class-specific complexities, and refine the model or dataset accordingly (Grandini et al. 2020).

3.5.3 Feature importance

After training the classifiers to categorize the four cities, a feature importance analysis is carried out using the Permutation Feature Importance technique (Kaneko 2023). This method systematically randomizes each feature, breaking its link to the city label and assessing how much it affects the model's performance, measured here by the AUC. Features that cause the most significant drop in AUC upon permutation are interpreted as making the most significant contribution to the classification outcome. In practical terms, the approach permutes each feature multiple times and records the average and variability of the AUC changes. These results are aggregated into a ranked list of features, clearly indicating which inputs—whether text-based or image-based—strongly influence the classification. By highlighting elements such as specific emotion-driven attributes from DistilBERT or compositional properties derived from CLIP, it becomes evident which factors are pivotal for distinguishing between the four cities. The top five features were selected because their removal produced a marked decrease in AUC, whereas features ranked lower had only marginal effects on performance. This threshold ensures that only dimensions with significant predictive power are retained for analysis.

3.6 Text-generated and image-generated feature correlation matching study

Building upon the results of the previous analytical steps, the final stage involves exploring how the top five DistilBERT features—derived from emotion recognition—and the top five CLIP features—about visual attributes—may correlate with each other (James et al. 2013). The central aim is to identify how particular emotional states (e.g., anger, joy, or admiration) align with specific visual features (e.g., brightness, contrast, or sharpness), thereby offering insight into whether specific visual patterns evoke or transmit corresponding affective responses (Zhang et al. 2021). This correspondence analysis relies on correlation metrics (such as Pearson's R), with each of the five DistilBERT-derived emotion features compared pairwise against each of the five CLIP-derived visual features in a correlation matrix. High positive or negative coefficients suggest meaningful relationships that could deepen the interpretability of generative image outputs about the textual emotional content.

We also perform a classification experiment, which is conducted using a combined 10-feature set consisting of the five most influential DistilBERT emotions and the five most salient CLIP attributes. By integrating these features into a single model, the study evaluates whether multi-modal synergy—linking emotional and visual cues—enhances predictive accuracy over classifications relying solely on text-based or image-based embeddings.

4 Results and discussion

4.1 Subjective characteristics and analysis

Based on a meticulous, manual review of all collected texts, the table reflects a comprehensive yet distilled representation of shared attributes across Berlin, Dublin, Málaga, and Cairo. Throughout this process, particular attention was given to removing direct identifiers—such as specific district names or iconic landmarks—thereby minimizing potential biases that could inadvertently influence classification or sentiment analysis. By scrutinizing each piece of content and categorizing common threads, researchers isolated each city's most salient themes related to cultural, historical, architectural, and environmental factors.

This subjective reading also highlighted transversal topics, such as urban ambiance, culinary distinctions, and climatic variations, which emerge consistently across all four destinations. The result is a high-level synthesis of user perceptions that preserves essential nuances of travelers' experiences (Table 2). Categories like "Historical Heritage" and "Color & Light" thus capture overarching impressions without violating anonymization standards.

4.2 Image data generation

Figure 4 displays four representative images, each generated according to the guidelines outlined in the methodology section. Although these visuals have been created to capture the essence of tourist experiences in Berlin, Dublin, Málaga, and Cairo,

Table 2 Subjective cross-city comparative characteristics

Characteristic	Berlin	Dublin	Málaga	Cairo
Historical Heritage	Marked by 20th-century events and monuments, evidence of wartime and post-reunification evolution.	Influenced by Viking and Celtic legacies, with medieval traces and literary landmarks.	Reflects layers of Phoenician, Roman, and Moorish histories in forts and old town areas.	Home to millennia-old monuments (e.g., pyramids) showcasing ancient and Islamic heritage.
Cultural & Artistic Scene	Global museums, street art, electronic music, and diverse festivals.	Literary traditions, pub music, and strong local identity; frequent live performances.	Rich in museums and festivals, blending flamenco and modern nightlife.	Islamic art, bustling markets, and modern cultural venues coexist with historical mosques.
Climate & Environment	Mild summers, cool winters, and plentiful parks; light often subdued.	Characterized by cloudy, rainy weather and lush greenery.	Sunny Mediterranean climate with warm beaches and bright natural light.	Hot and arid, but irrigated by the Nile; intense sunshine year-round.
Architecture	Combines historical landmarks, modern innovation, and alternative spaces.	Georgian facades, Gothic cathedrals, and cobblestone streets; some contemporary infill.	Moorish influence, ancient fortifications, and contemporary structures.	Ancient wonders, Islamic architecture, and modern developments in a sprawling metropolis.
Lifestyle & Social Atmosphere	Cosmopolitan with an intense nightlife scene; eclectic, multicultural, and forward-thinking.	Warm, communal pub culture, convivial gatherings, and storied traditions.	Festive and coastal, with a welcoming spirit emphasizing outdoor activities and tourism.	Dynamic and bustling, centered on communal gatherings and ever-active markets.
Culinary & Gastronomy	Rich street-food culture, breweries, and diverse international cuisines.	Renowned beer culture, hearty pub fare, and evolving modern Irish cuisine.	Famed for seafood, tapas, and Mediterranean flavors, strong café culture.	Middle Eastern dishes featuring aromatic spices, street fare, and iconic staples (e.g., kosher, falafel).
Color & Light	Grays, greens, neon graffiti, diffused light in cooler seasons.	Dominated by gray skies and green terrain, pops of color from doors and signboards.	Warm oranges, blues, and vivid sunlight reflect a lively coastal environment.	Ochre and golden hues under a bright sun; desert ambiance near outlying areas.
Distinctive Aromas	Urban mix of food stalls and greenery, occasionally beer undertones.	Damp cobblestones, malt, and beer scents around traditional pubs.	Fresh sea air, citrus blossoms, and grilled fish along the coast.	Fragrant spices, street-food aromas, and desert dust, especially in peripheral zones.

they remain sufficiently abstract to avoid explicit references to specific landmarks. Each image nevertheless conveys certain defining attributes that align with the cultural or environmental context of the city in question. For example, portraying a prominent waterway can evoke historical or geographical themes without overtly identifying the location. These images illustrate how generative models can produce

Fig. 4 Instances of image generation. Berlin (upper left), Dublin (upper right), Málaga (lower left), Cairo (lower right)

Table 3 Models and performance metrics classification after feature extraction with fasttext from reviews

Model	AUC	CA	F1	Prec	Recall
LR	0.976	0.860	0.860	0.864	0.860
NN	0.978	0.880	0.878	0.884	0.880
SVM	0.991	0.930	0.930	0.933	0.930

Table 4 Confusion matrix with the best model and feature extraction with fasttext from reviews

		Predicted			
		BERLÍN	DUBLÍN	EI CAIRO	MÁLAGA
Actual	BERLÍN	85.7%	0.0%	4.0%	0.0%
	DUBLÍN	0.0%	100.0%	0.0%	4.3%
	EI CAIRO	7.1%	0.0%	92.0%	0.0%
	MÁLAGA	7.1%	0.0%	4.0%	95.7%

visual summaries that reflect common perceptions of a city's architecture, climate, and lifestyle while respecting anonymity and minimizing bias.

4.3 Embedding methods. Classifications and performance

Once the tourist review corpus and the corresponding DALL-E-generated images are established, latent feature representations will be extracted through the described embedding methods, FastText and Inception V3. These high-dimensional, non-interpretable embeddings will then be utilized to train and evaluate the aforementioned classifiers, aiming to estimate the maximum classification metrics and further validate the robustness of the proposed multi-modal framework. Tables 3 and 4 show the metrics and confusion matrix for the best model in the case of classification with

Table 5 Models and performance metrics classification after feature extraction with inception V3 from images

Model	AUC	CA	F1	Prec	Recall
LR	0.948	0.810	0.809	0.820	0.810
NN	0.925	0.810	0.811	0.813	0.810
SVM	0.970	0.850	0.852	0.858	0.850

Table 6 Confusion matrix with the best model and feature extraction with inception V3 from images

		Predicted			
		BERLÍN	DUBLÍN	EI CAIRO	MÁLAGA
Actual	BERLÍN	79.2%	4.2%	10.0%	9.1%
	DUBLÍN	8.3%	91.7%	0.0%	4.5%
	EI CAIRO	0.0%	0.0%	83.3%	0.0%
	MÁLAGA	12.5%	4.2%	6.7%	86.4%

latent features extracted for reviews, and Tables 5 and 6 reciprocally show results with features extracted for images. With the classification based on embedding methods, we are confident that there are highly discriminating structures among the four categories.

Initial experiments suggest that high-dimensional embeddings from FastText and Inception V3 already encapsulate salient information pertinent to city differentiation, leading to high classification performance regarding AUC, Accuracy, and F1-scores. As a result, while these embeddings excel at pattern recognition, researchers and stakeholders may desire additional explainable methods to justify classification decisions (Wang et al. 2024).

It should be noted that although the performance metrics are high across all assessed methods, the image-based pipeline exhibits slightly lower values due to the informational loss incurred when transforming textual content into generated imagery. Nonetheless, the fact that all three classifiers report comparably strong metrics signals a robust underlying structure within the data, indicative of pronounced distinctions among the tourist destinations. Consequently, it remains feasible to classify these locations reliably, and the resulting classification outcomes effectively offer an upper bound on the achievable level of discrimination under this methodological framework.

4.4 Explainable methods. Classifications and performance

4.4.1 Classification of destinations with explainable features obtained from reviews with DistilBERT

Using DistilBERT for feature extraction by semantic identification from reviews, it can be seen in Tables 7 and 8 that we obtain somewhat lower but equally good metrics with a balanced distribution in the confusion matrix. It can be observed that very acceptable results are obtained without reaching the maximum reference values of metrics of the embedding method FastText (Tables 3 and 4).

Furthermore, a permutation test on the top-performing classifier confirms that these findings are not attributable to chance (Fig. 5). Specifically, as the proportion of

Table 7 Models and performance metrics classification after feature extraction with DistilBERT from reviews

Model	AUC	CA	F1	Prec	Recall
LR	0.943	0.780	0.779	0.781	0.780
NN	0.945	0.790	0.789	0.790	0.790
SVM	0.944	0.810	0.809	0.816	0.810

Table 8 Confusion matrix with the best model and feature extraction with DistilBERT from reviews

		Predicted			
		BERLÍN	DUBLÍN	EI CAIRO	MÁLAGA
Actual	BERLÍN	78.3 %	15.4 %	13.0 %	0.0 %
	DUBLÍN	13.0 %	73.1 %	0.0 %	10.7 %
	EI CAIRO	8.7 %	7.7 %	82.6 %	7.1 %
	MÁLAGA	0.0 %	3.8 %	4.3 %	82.1 %

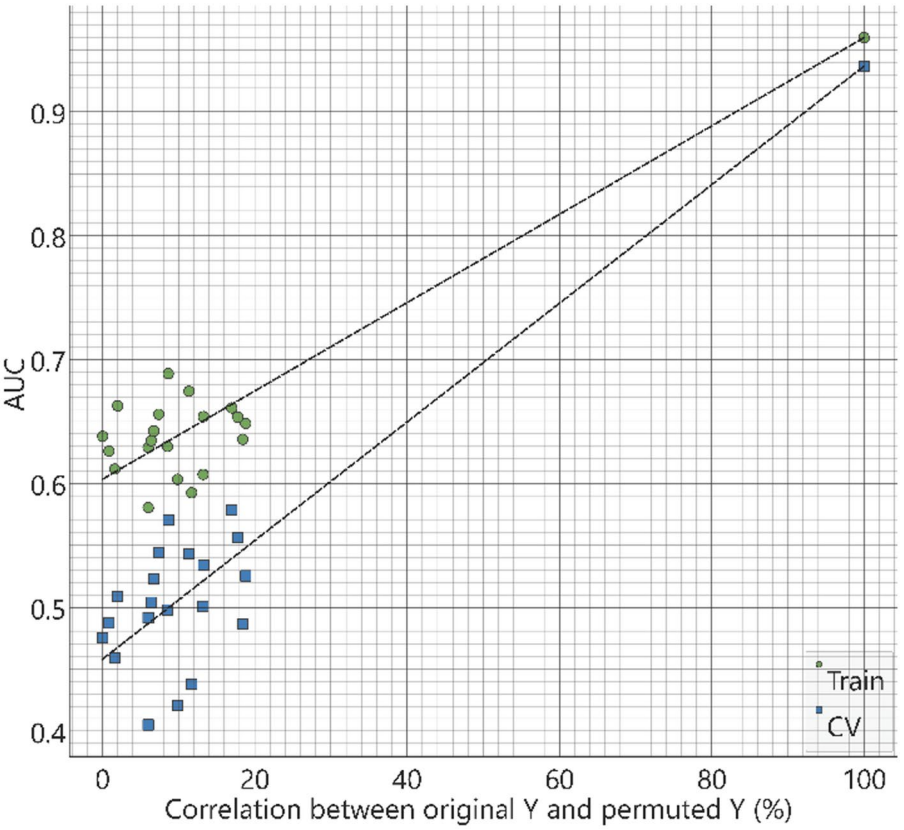

Fig. 5 Permutation plot of the best model and feature extraction with DistilBERT from reviews

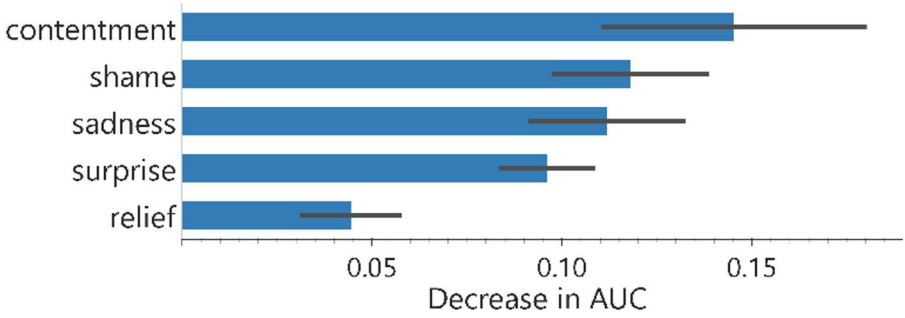

Fig. 6 Feature importance using the best model and feature extraction with DistilBERT from reviews

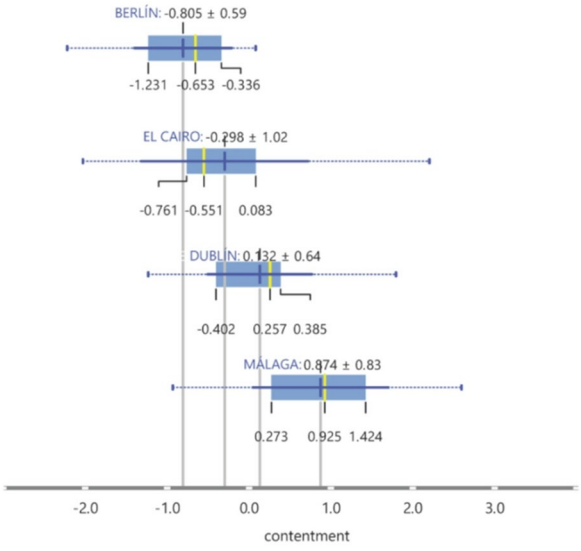

Fig. 7 Boxplot distribution of 'contentment' emotion

correctly labeled (non-permuted) instances decreases, the AUC also declines, a pattern evident in both single-phase training and cross-validation outcomes.

Among the emotions analyzed, **contentment**, **shame**, **sadness**, **surprise**, and **relief** emerge as the most discriminating factors for classifying tourist reviews between Berlin, Dublin, Cairo, and Málaga (Fig. 6). This distinction stems from each city's unique cultural, climatic, and historical elements, eliciting different emotional responses (Rasul et al. 2024).

Contentment is a powerful differentiator because it aligns with how travelers perceive a harmonious balance between a destination's cultural offerings and personal comfort in subjects like customer service, accessibility, or social atmosphere (Gupta et al. 2024). In cosmopolitan settings known for efficiency and rich social scenes, visitors often express contentment when basic needs—such as ease of transportation or favorable climate—are met with stimulating yet not overwhelming urban environments. Contentment emerges as a significant discriminator when examining the boxplot for the four destinations (Fig. 7), revealing how travelers' overall sense of

satisfaction diverges between locales. The chart demonstrates that Berlin exhibits a markedly lower median contentment level than the other three cities (-0.805 ± 0.59), potentially reflecting particular challenges visitors encounter—ranging from historical trauma and inclement weather remnants to perceptions of a more formal social atmosphere. By contrast, El Cairo achieves better contentment scores(-0.298 ± 1.02), suggesting that while tourists appreciate the city's profound cultural heritage and hospitable communities, their comfort might be tempered by the demands of navigating its dense urban environment or coping with extreme temperatures (Athukorala and Murayama 2021). Notably, both Dublin (0.132 ± 0.64) and Málaga (0.874 ± 0.83) display comparatively higher scores, indicating that the general tourist experience tends toward a smoother balance between practical amenities (e.g., straightforward communication, accessible transportation) and enjoyable social or leisure offerings. Although Dublin's variable climate can pose a minor hindrance, its storied literary scene, lively pubs, and convivial public life help foster a sense of ease and fulfillment. Málaga, benefiting from a relaxed Mediterranean ambiance and scenic waterfronts, often provides a favorable climate year-round, positively influencing contentment levels.

Shame, in contrast, frequently arises when visitors confront dissonant or unsettling aspects of a location's history, societal inequalities, or cultural norms. As the boxplot depicts, it emerges as a distinctly polarized emotion across these four destinations (Fig. 8). El Cairo, with a notably higher mean shame level (0.665 ± 0.69), suggests that travelers often confront societal or infrastructural disparities that trigger moral unease (Nawijn and Biran 2019). They may also experience culture shock when encountering the city's rapid modernization juxtaposed against longstanding traditions and economic imbalances. Dublin shows a positive, albeit lower, shame average (0.460 ± 0.71), potentially reflecting visitors' introspections about the city's complex past and persisting social narratives, from historical conflicts to religious

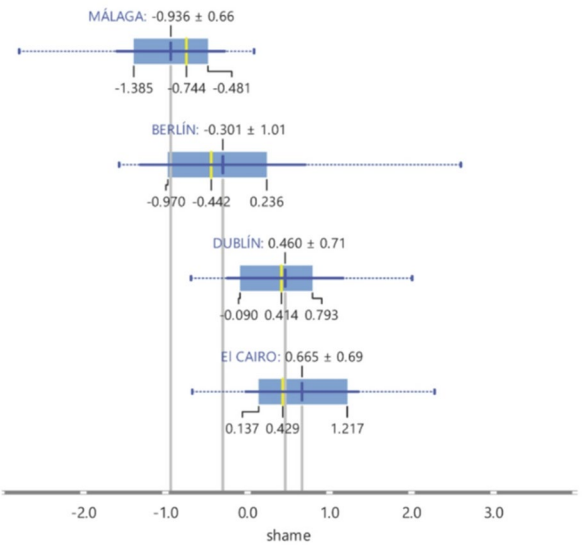

Fig. 8 Boxplot distribution of 'shame' emotion

influences. By contrast, Berlin appears near or below zero on average (-0.301 ± 1.01), indicating that although its layered wartime legacy can spark reflection, visitors may be more inclined to regard the city's commemorative efforts and modern reinvention with respect rather than guilt. Lastly, Málaga exhibits the lowest levels of shame (-0.936 ± 0.66), consistent with the region's generally relaxed atmosphere and the relative absence of overt socio-political tension in the tourist gaze. These varied intensities of shame thus reveal how travelers' ethical awareness or discomfort is shaped by each locale's historical memory, public discourse, and visible societal contexts. Reviews that articulate such affective dissonance underscore deeper engagement with a destination's social reality, providing a strong basis for distinguishing among Berlin, Dublin, El Cairo, and Málaga.

Sadness often mirrors a traveler's emotional response to perceived gloom—whether from climate, urban melancholy, or the weight of historical memory—and thus emerges with varying intensity across diverse locales (Jordan et al. 2019). In cities prone to overcast or wet weather, like Dublin, visitors might voice sadness concerning fleeting daylight hours or chilly, damp conditions that temper explorations. Conversely, cities with strong memorial cultures or striking vestiges of conflict, such as Berlin, may evoke a more historical or contemplative sadness tied to reminders of past events. In both scenarios, the traveler's sentiment echoes the local environment in a way that distinctly demarcates one place from the next, facilitating robust discrimination among destination reviews.

Surprise is closely tied to how a location diverges from a visitor's prior assumptions or experiences, making it a strong classifier for places with vivid contrasts or abrupt shifts in cultural rhythms (Mehra 2023). For instance, Cairo's juxtaposition of modern and ancient elements can deliver an intense sense of wonder, just as Dublin's sudden shifts in weather or impromptu pub sessions defy expectations for travelers unaccustomed to such spontaneous sociocultural phenomena. Where travelers note genuine astonishment at architectural style, local customs, or even the flavor profiles of traditional cuisine, that sense of "surprise" tends to be unique to each urban context. This emotional marker thus provides a experience-based indicator of the destination's ability to defy familiar norms.

Relief, conversely, encapsulates the moment travelers find respite from stressors, whether navigating unfamiliar transit systems in Berlin, transitioning from arid midday heat to more extraordinary interior spaces in Cairo or adapting to sudden meteorological changes in Dublin (Park 2024). This feeling emerges in reviews as a profoundly personal acknowledgment of overcoming perceived hardships. The specificity of what triggers relief—the ease of finding English signage, the comfort of a shaded café, or the resolution of initial language barriers—varies across locales. As a result, relief supplies a telling emotional cue, signaling the kinds of tension each city introduces and how readily that tension is resolved.

Taken together, contentment, shame, sadness, surprise, and relief offer a layered emotional spectrum that can be reliably linked to the distinct sociocultural, climatic, and historical features characterizing Berlin, Dublin, Cairo, and Málaga, not only delineating individual travel experiences but also yielding structured, discriminating signals that can be exploited to classify reviews with high accuracy.

Table 9 Models and performance metrics classification after feature extraction with CLIP from images

Model	AUC	CA	F1	Prec	Recall
LR	0.871	0.670	0.670	0.675	0.670
NN	0.866	0.630	0.632	0.638	0.630
SVM	0.839	0.600	0.601	0.604	0.600

Table 10 Confusion matrix with the best model and feature extraction with CLIP from images

		Predicted			
		BERLÍN	DUBLÍN	El CAIRO	MÁLAGA
	BERLÍN	72.0%	19.0%	12.0%	0.0%
	DUBLÍN	4.0%	71.4%	8.0%	24.1%
Actual	El CAIRO	8.0%	4.8%	68.0%	17.2%
	MÁLAGA	16.0%	4.8%	12.0%	58.6%

4.4.2 Classification of destinations with explainable features obtained from images with CLIP

CLIP is used to extract features by semantic identification from images, with which we obtain somewhat lower metrics than with DistilBERT but equally adequate (Tables 9 and 10). The decrease in performance is due to the loss of information due to the text-image transformation of the data. Although there is now an apparent decrease concerning the equivalent embedding method (Inception V3, Tables 5 and 6), the gain in interpretability does not reduce the classification performance so markedly.

A permutation test applied to the most effective classifier provides additional evidence that these results are not due to chance (Fig. 9). As fewer instances remain correctly labeled (i.e., not permuted), the AUC consistently diminishes—a pattern in single-phase training and cross-validation assessments.

In the context of AI-generated images derived from tourist reviews, the five visual features of **balance**, **brightness**, **entropy**, **dynamic range**, and **colorfulness**, after a process of feature ranking, stand out as particularly effective for distinguishing among Berlin, Dublin, Cairo, and Málaga due to the distinct cultural, climatic, and aesthetic impressions each city conveys (Fig. 10).

In photographic editing, "balance" often refers to the harmony of color temperature and tint adjustments—collectively known as white balance—necessary to ensure that scene colors appear natural. A higher positive value on the boxplot typically corresponds to a warmer, more yellowish, or reddish cast, whereas a more negative value suggests a cooler, bluer tint (Yu et al. 2023). Examining the four destinations through this lens reveals noticeable differences in their median and spread of white-balance values (Fig. 11), hinting at contrasting ambient lighting conditions or predominant color casts inferred from user-generated content. El Cairo, showing a noticeably high mean and median (around 0.996 ± 0.74), suggests that the AI-generated images for this city frequently adopt warmer tones, consistent with desert sun and golden-hued

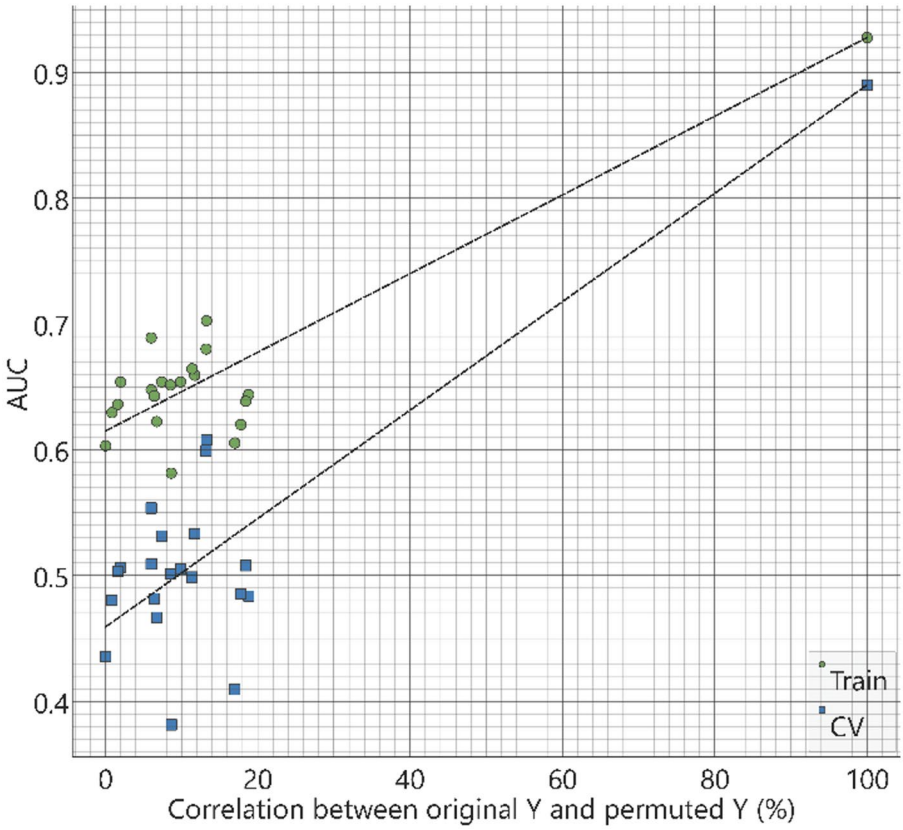

Fig. 9 Permutation plot of best model and feature extraction with CLIP from images

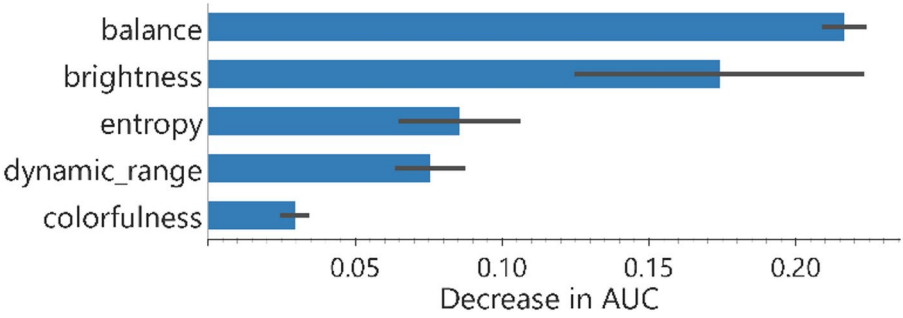

Fig. 10 Feature importance using the best model and feature extraction with CLIP from images

architecture. In contrast, Dublin exhibits markedly lower (negative) balance values (around -0.573 ± 0.79), aligning with cooler, overcast, or more temperate lighting conditions often described in reviews of its rainy maritime climate. Berlin lies at -0.320 ± 0.82, indicative of milder but still somewhat cool undertones in the images,

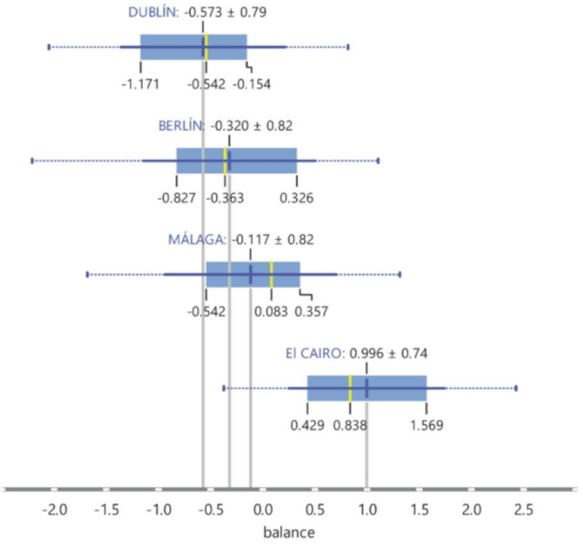

Fig. 11 Boxplot distribution of 'balance' feature

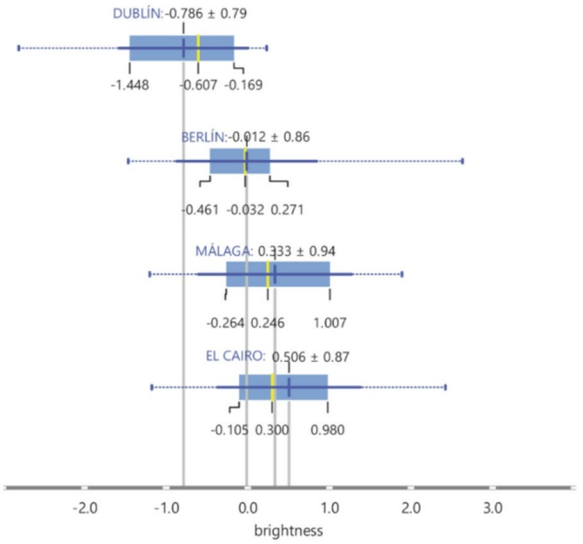

Fig. 12 Boxplot distribution of 'brightness' feature

perhaps evoking its mix of modern steel-and-glass structures and temperate-zone weather. Meanwhile, Málaga is close to neutrality (-0.117 ± 0.82), reflecting an often sunny environment but not to the same degree of warmth found in desert climates, leading to more moderate color temperatures.

Brightness, meanwhile, is critical for reflecting the climate and environmental factors embedded in user reviews, as travelers often comment on the quality of light or overall luminance of a locale. In the boxplot (Fig. 12), El Cairo stands out for having the highest average brightness (0.506 ± 0.87), which aligns with textual prompts mentioning intense desert sun, reflective golden landscapes, or even warmly lit indoor

markets (Ship et al. 2024). Málaga, with an average brightness of 0.333 ± 0.94, similarly evokes images of a sunny coastal environment, albeit less consistently intense than Cairo's desert climate. In contrast, Berlin hovers close to neutral at -0.012 ± 0.86, reflecting references to variable weather and a blend of open urban spaces, historic neighborhoods, and modern architecture. At the same time, many visitors may mention cloudy days or stark winter light, while others highlight bright summer festivals and bustling outdoor markets. Finally, Dublin presents the lowest brightness value at -0.786 ± 0.79, corroborating perceptions of overcast skies, frequent rainfall, and generally subdued lighting conditions. Although the textual content of user reviews ultimately shapes these AI-generated images, the boxplot suggests that collectively, travelers perceive and describe each city's ambiance in ways that systematically translate into distinct brightness levels.

Entropy captures the degree of apparent disorder or complexity within an image, determined by how many distinct patterns, shapes, or textures compete for the viewer's attention (Zhang et al. 2020). Cairo's bustling markets and richly ornamented mosques might yield scenes marked by an elevated entropy. In contrast, although busy in certain locales, Dublin's cityscapes may exhibit a comparatively moderate level of visual complexity. In Berlin, subcultural street art and avant-garde architectural experimentation can spike entropy levels, but certain administrative and cultural districts remain understated and streamlined. Málaga's coastal vistas, by contrast, might present more uniform color zones of sea and sky, reducing entropy except in the crowded urban center.

Dynamic range evaluates the span from darkest shadows to brightest highlights, reflecting how photographers—and, by extension, AI models—represent strong sunlight, deep shadows, or nighttime illuminations (Yu et al. 2023). High dynamic range is typical of sun-intense locales, with Cairo's stark differences between bright desert sun and covered market corridors being a quintessential example. Málaga's coastal sunscapes, too, can lead to richly contrasted scenarios, capturing the glow of sunlight on the water while retaining detail in shaded archways. In a city like Dublin, variable weather patterns often generate more diffused lighting, thus moderating the dynamic range in user-generated or AI-synthesized images. Berlin, with a mix of historical monuments and futuristic structures, can vary widely depending on the portrayal of interiors, nighttime city lights, or open-air vistas. Because travelers often incorporate comments about vivid sunlight or striking illumination into their texts, dynamic range emerges as a salient discriminant when rendering images.

Colorfulness, finally, underscores the variety and saturation of hues depicted in an AI-generated scene. Málaga's Mediterranean palette—dominated by vibrant blues of the sea and bright yellows of southern European facades—contrasts sharply with Dublin's more subdued color spectrum of stone architecture and frequent grays of overcast skies (Li et al. 2024). Cairo's palette, often characterized by sandy browns and warm earth tones, can reveal pockets of bold color in ornate textiles or lively market stalls, pushing colorfulness upward in specific contexts. Berlin's street art culture provides another vivid source of color infusion, although austere neighborhoods or minimalist modern architecture may reduce color variety in other parts of the city. Because these cities vary in climate, historical development, and architectural practice, the resultant color palettes differ significantly enough to register as

class-distinguishing features once the user reviews prompting the image creation have described relevant scenery.

Altogether, these five features—balance, brightness, entropy, dynamic range, and colorfulness—reflect the cumulative impact of environmental lighting, architectural arrangement, cultural context, and traveler perception. The synergy of these attributes within AI-generated images enables robust discrimination between Berlin, Dublin, Cairo, and Málaga, effectively capturing the traveler's impressionistic worldview encoded in their reviews, and provides a data-driven approach to understanding how visitors visually conceptualize and differentiate among these diverse cities.

4.5 Feature vector representations

One of the most challenging aspects of high-dimensional feature representations is that they are not straightforward to visualize or interpret. To address this, we applied t-SNE, a non-linear dimensionality-reduction method, to project our extracted feature vectors from multiple models (FastText, Inception V3, DistilBERT, and CLIP) into a two-dimensional space. As shown in Fig. 13, the resulting plots highlight how different classes (cities) tend to form distinct clusters while still revealing points of overlap that mirror the observed confusion in classification. This visualization helps

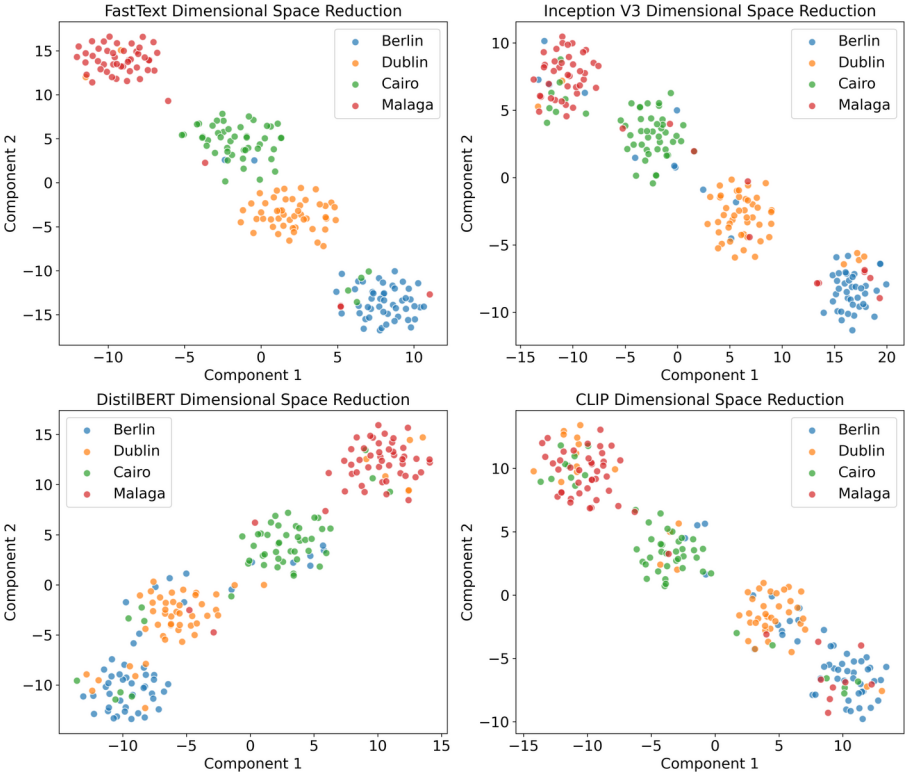

Fig. 13 Dimensional Space Reductions with t-SNE

Table 11 Models and performance metrics classification combining textual and visual features

Model	AUC	CA	F1	Prec	Recall
LR	0.967	0.850	0.850	0.850	0.800
NN	0.967	0.840	0.839	0.840	0.787
SVM	0.973	0.840	0.843	0.859	0.791

Table 12 Confusion matrix with the best model combining textual and visual features

		Predicted			
		BERLÍN	DUBLÍN	EI CAIRO	MÁLAGA
Actual	BERLÍN	84.0%	8.0%	8.0%	0.0%
	DUBLÍN	16.0%	80.0%	0.0%	4.0%
	EI CAIRO	0.0%	12.0%	84.0%	4.0%
	MÁLAGA	0.0%	0.0%	8.0%	92.0%

illustrate the degree of separation, the underlying relationships among the feature vectors, and the overall discriminative power of the representations learned by our approach.

4.6 Explainable cross-modal interactions: textual emotions and visual attributes

4.6.1 Textual and visual features in machine learning classification performance

By selecting the top five DistilBERT-derived emotion features—contentment, shame, sadness, surprise, and relief—and combining them with five key visual attributes derived from AI-generated images—balance, brightness, entropy, dynamic_range, and colorfulness—, we obtain a lean yet potent feature set that substantially enhances classification performance over using either all text-based emotions or all image-based features alone (Priya and Udayan 2020). This improvement is reflected in higher overall metrics (Table 11) and a more cohesive confusion matrix whose off-diagonal entries decrease (Table 12), indicating fewer misclassifications across the four city classes. The metrics obtained are higher than those obtained with the separately explainable methods and are in the middle of the embedding methods analyzed (FastText and Inception V3, Tables 3, 4, 5 and 6). The synergy arises because the selected emotions capture distinct affective nuances tied to travelers' subjective experiences, while the visual properties encode the destinations' complementary compositional and aesthetic aspects. These attributes form a well-rounded, semantically rich, and discriminative representation, reflecting each city's unique identity from dual perspectives—human emotional response and AI-interpreted visual motifs. The integrated feature set effectively leverages meaningful contrasts in climate, cultural heritage, and social environments, thereby reducing noise and enhancing the model's ability to partition user reviews with greater precision. This outcome reinforces the potential of blending interpretable, sentiment-driven dimensions with targeted image-based metrics for robust, explainable multi-modal classification in tourism analytics.

Figure 14 underscores how a carefully chosen blend of textual and visual features yields greater classification power than either modality alone. This relationship suggests that neither emotional nuances nor compositional cues fully capture the com-

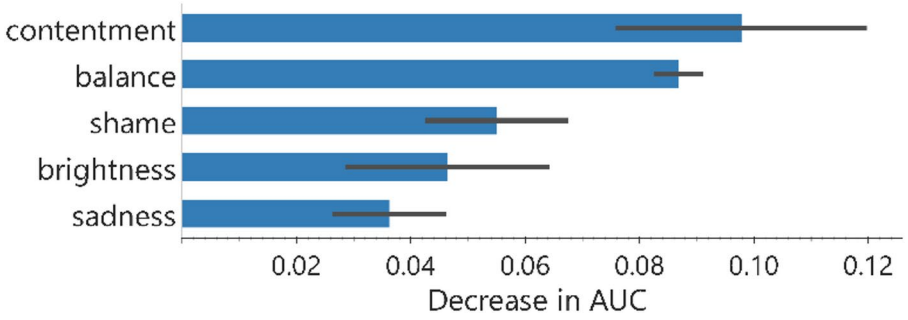

Fig. 14 Feature importance using the best model combining textual and visual features

Table 13 Visual and emotional features correlations

Visual feature	Emotion feature	Correlation
brightness	contentment	+0.646
balance	relief	+0.604
entropy	shame	+0.544

plexity of a destination on its own. Instead, the synergy between them better reflects the multilayered experiences travelers describe, leading to a deeper understanding—and more accurate classification—of each location's unique identity (Al-Tameemi et al. 2023).

4.6.2 Correlation analysis

To deepen our understanding of how textual emotions and visual characteristics jointly shape tourists' perceptions, we will conduct a correlation analysis between the sentiment scores and the selected image-based features using CLIP. Correlation coefficients will be interpreted following Cohen's (1988) guidelines, where values around 0.1 indicate a small effect size, around 0.3 suggest a medium relationship and 0.5 or higher denotes a strong association. By identifying only strong correlations (Table 13), we can infer how specific visual properties align with particular affective states. Ultimately, this mapping could inform strategic choices in promotional imagery, enabling practitioners to craft visuals more likely to evoke positive emotions among prospective travelers. In the longer term, correlational insights might pave the way for automated systems that adapt images to specific emotional tones, thus enhancing the resonance and impact of tourism marketing campaigns.

The high positive correlation between brightness and contentment ($r = +0.646$) can be understood through theoretical and empirical findings in environmental psychology and neuroaesthetics (Fig. 15). First, extensive research has demonstrated that luminance is pivotal in modulating affective responses. Valdez and Mehrabian (1994) found that increased brightness is generally associated with higher arousal and more positive mood ratings, suggesting that brighter environments evoke feelings of well-being and contentment. Similarly, Nixon et al. (2023) reported that higher ambient light levels—especially indoors—enhance mood states by creating an energizing and emotionally uplifting atmosphere. This body of work supports the

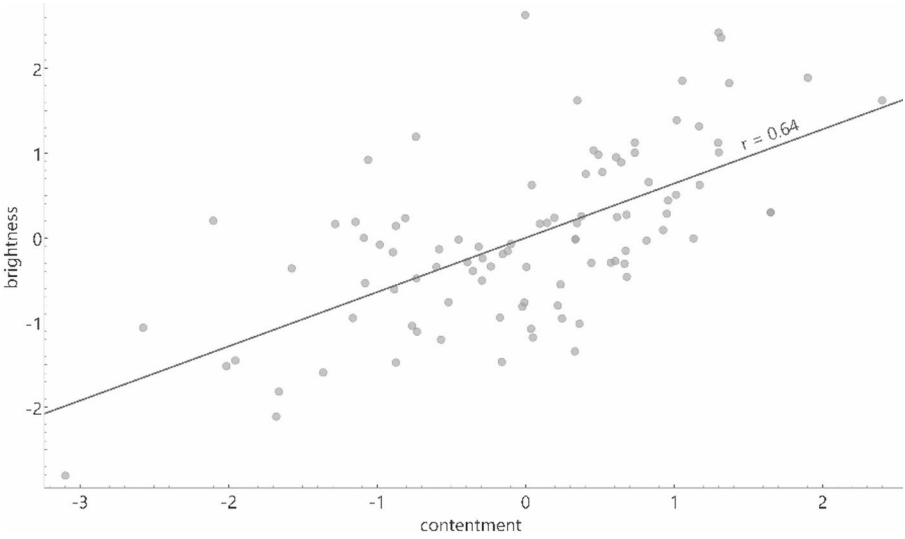

Fig. 15 Correlation' brightness' vs. 'contentment'

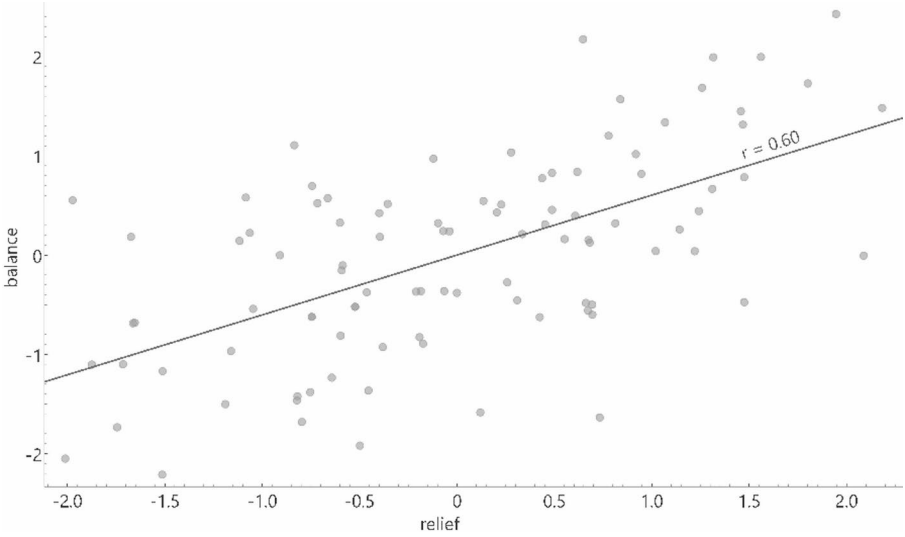

Fig. 16 Correlation 'balance' vs. 'relief'

notion that brightness is a proxy for favorable environmental conditions, facilitating positive emotional outcomes.

The strong positive correlation between balance and relief ($r \approx +0.604$) can be explicated by integrating findings from visual perception and affective neuroscience (Fig. 16). Empirical studies have shown that when images are rendered with a white balance that closely mimics natural daylight, the resulting visual stimulus is more aesthetically pleasing and reduces perceptual dissonance. This phenomenon has been

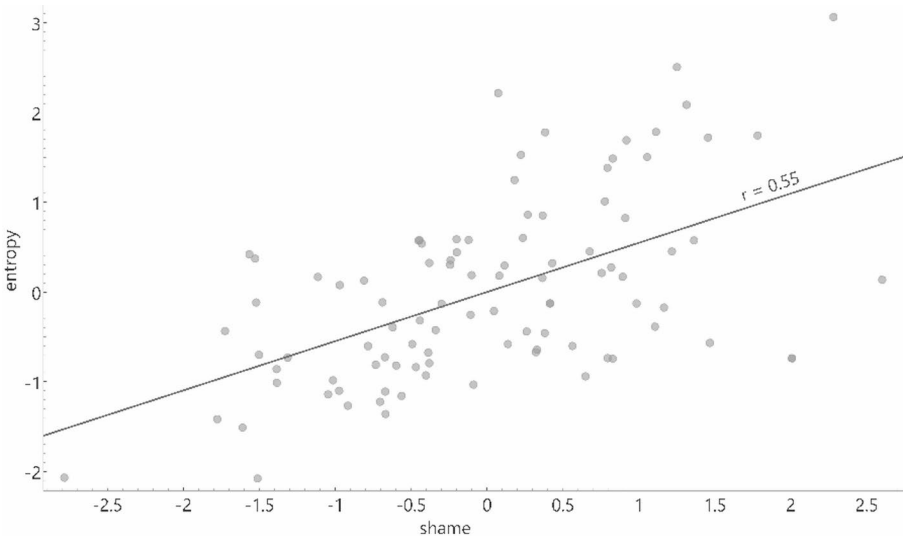

Fig. 17 Correlation 'entropy' vs. 'shame'

linked to lower cognitive load and enhanced viewer comfort, thereby eliciting feelings of relief (Nixon et al. 2023). Brainard & Wandell's (1986) work on color constancy underscores that the human visual system is optimized for interpreting scenes under typical environmental lighting; deviations from this norm can heighten perceptual effort. Conversely, an optimal balance minimizes such discrepancies, enabling a more fluent perceptual experience that is inherently soothing. This reduced cognitive demand likely contributes to a subjective sense of relief, as the observer is freed from the strain associated with correcting or mentally "recoloring" the scene.

The significant positive correlation between entropy and shame ($r \approx +0.544$) is shown in Fig. 17. From a psychological standpoint, perceptual disfluency has been shown to elicit negative affective responses. When an image is characterized by high entropy, the resultant perceptual incoherence can lead to an unpleasant experience that may be subconsciously interpreted as a departure from normative aesthetic order. This disruption is particularly salient when considering self-conscious emotions such as shame. Shame is not only an adverse effect but also a socio-cognitive response that arises from the perception of having deviated from culturally or personally internalized standards of order and propriety (Laing 2021). In this light, high entropy in a visual stimulus may metaphorically mirror internal disarray or inadequacy, thereby potentiating feelings of shame.

Our analysis focuses on broad, general correlations between text-derived emotional cues and image-based visual features across all cities without delving into urban differences. While we recognize that urban attributes likely influence these relationships, a detailed examination of urban differences would require segmentation of the dataset by city and including supplementary variables capturing urban context.

5 Conclusions and future work

In this study, we introduced an innovative multi-modal framework that transcends traditional sentiment and topic modeling approaches in tourism analytics by integrating textual and visual data in a novel manner. Our approach pioneers the transformation of user-generated textual reviews into AI-generated images using a standardized prompt, thereby enabling the conversion of latent emotional cues into explicit visual features. This transformation facilitates mapping affective signals—extracted via advanced models such as DistilBERT—onto interpretable visual attributes derived from CLIP. The resulting framework captures the subtle emotional states expressed in tourist reviews and harnesses these affective insights to generate standardized, language-agnostic imagery. Such imagery, in turn, supports the extraction of explainable features, such as brightness, entropy, and white balance, which reflect the underlying structure of tourist perceptions across diverse destinations.

Our experimental results underscore the viability of this integrated methodology, demonstrating high classification accuracy when distinguishing between cities with markedly different cultural, climatic, and historical backgrounds. The ability to correlate text-derived sentiments with specific visual attributes provides compelling evidence of an inherent structure within tourist perceptions, evaluated at maximum discriminative performance by embedding methods (FastText and Inception V3). This structure enables robust classification and offers actionable understandings for destination managers and policymakers when effectively characterized through explainable features. Notably, the fusion of sentiment-driven dimensions with visual features represents a significant advancement over conventional "black box" models, offering a more transparent and subtle understanding of how tourist experiences are visually and emotionally constructed. The novelty of our work lies in several key contributions. First, transforming text to image is a powerful tool to homogenize diverse textual data into a visually consistent format, mitigating issues related to language and cultural disparities. Second, by converting sentiments into tangible visual features, we narrow the gap between abstract emotional expressions and their concrete aesthetic manifestations. Third, our framework leverages explainable feature extraction to uncover underlying structures in tourist perceptions, thereby facilitating enhanced classification and interpretability. Finally, integrating multi-modal data—through a seamless combination of text and image features—opens new avenues for exploiting these insights in tourism management, urban planning, and marketing strategies.

Despite these promising findings, several avenues for future research remain. First, further investigation is warranted into the scalability and generalizability of the proposed framework across a broader range of destinations and languages. Expanding the dataset to include a more diverse set of tourist reviews, especially from non-English sources, could enhance the robustness of our findings. Second, future work should explore incorporating real-time data streams to dynamically update sentiment and visual feature correlations, enabling adaptive marketing strategies and responsive urban development initiatives. Third, additional research is needed to refine the transformation process from text to image, perhaps by integrating state-of-the-art generative models that can capture even subtler nuances of traveler sentiment.

Finally, the potential integration of this multi-modal approach with geospatial and temporal analytics could yield a more comprehensive understanding of how tourist perceptions evolve over time and across different geographic contexts.

Author contributions Conceptualisation, V.C.-F.; methodology, V.C.-F. and I.R.-R.; validation, V.C.-F., M.P.-C. and I.R.-R.; formal analysis, V.C.-F. and M.P.-C.; investigation, V.C.-F., M.P.-C. and I.R.-R.; resources, V.C.-F.; data curation, V.C.-F. and M.P.-C.; writing—original draft preparation, V.C.-F., M.P.-C. and I.R.-R.; writing—review and editing, V.C.-F., M.P.-C. and I.R.-R.; visualisation, V.C.-F. and M.P.-C.; supervision, I.R.-R. All authors have read and agreed to the published version of the manuscript.

Funding Funding for open access publishing: Universidad de Málaga/CBUA. This work is financed by by National Funds provided by FCT- Foundation for Science and Technology through project UIDB/04020: Centro de Investigação em Turismo, Sustentabilidade e Bem-Estar (Portugal); and by Grant RYC2023-045296-I funded by MICIU/AEI/10.13039/501100011033 and by ESF+, MICIU/AEI/10.13039/501100011033 (Spain).

Data availability The primary data used in this study were obtained from publicly available review platforms of tourist destinations.

Declarations

Competing interests The authors declare no competing interests.

References

Al-Tameemi I, Feizi-Derakhshi M, Pashazadeh S, Asadpour M (2023) Interpretable Multi-modal sentiment classification using deep Multi-View attentive network of image and text data. IEEE Access 11:91060–91081. https://doi.org/10.1109/ACCESS.2023.3307716

Alaei A, Becken S, Stantic B (2019) Sentiment analysis in tourism: capitalizing on big data. J Travel Res 58:175–191. https://doi.org/10.1177/0047287517747753

Athukorala D, Murayama Y (2021) Urban heat Island formation in greater cairo: Spatio-Temporal analysis of daytime and nighttime land surface temperatures along the Urban-Rural gradient. Remote Sens 13:1396. https://doi.org/10.3390/rs13071396

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Association Comput Linguistics 5:135–146. https://doi.org/10.1162/tacl_a_00051

Brainard DH, Wandell BA (1986) Analysis of the retinex theory of color vision. JOSA A 3(10):1651–1661

Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Kurzweil R (2018) Universal sentence encoder. arXiv preprint arXiv:1803.11175

Cilkin R, Çizel B (2022) Tourist gazes through photographs. J Vacation Mark 28:188–210. https://doi.org/10.1177/13567667211038955

Cohen J (1988) Set correlation and contingency tables. Appl Psychol Meas 12:425–434. https://doi.org/10.1177/014662168801200410

Cortes C, Vapnik V (1995) Support-Vector networks. Mach Learn 20:273–297. https://doi.org/10.1023/A:1022627411411
Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S (2020) GoEmotions: A dataset of Fine-Grained emotions. 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372
Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, 4171–4186. https://doi.org/10.18653/v1/N19-1423
Folino F, Ruga T, Zumpano E, Vocaturo E (2024) Visualizing tourism's future: the impact of image-based ai on destination development. 2024 IEEE Int Workshop Metrol Living Environ (MetroLivEnv) 81–86. https://doi.org/10.1109/MetroLivEnv60384.2024.10615477
Fu J, Zhou W, Jiang Q, Liu H, Zhai G (2024) Vision-Language consistency guided Multi-Modal prompt learning for blind AI-Generated image quality assessment. IEEE Signal Process Lett 31:1820–1824. https://doi.org/10.1109/LSP.2024.3420083
Gandelsman Y, Efros A, Steinhardt J (2023) Interpreting clip's image representation via Text-Based decomposition. ArXiv. https://doi.org/10.48550/arXiv.2310.05916.abs/2310.05916
Grandini M, Bagli E, Visani G (2020) Metrics for Multi-Class classification: an overview. ArXiv. abs/2008.05756
Gupta S, Kumar A, Ambulkar A, Kundra D, Venkadeshwaran K, Goyal S, Chakravarthy S (2024) Evaluating the Key Elements Contributing to Visitor Contentment at Cultural Festivals. Evolutionary Studies In Imaginative Culture. https://doi.org/10.70082/esiculture.vi.1138
Hao Y, Chi Z, Dong L, Wei F (2022) Optimizing prompts for Text-to-Image generation. ArXiv. https://doi.org/10.48550/arXiv.2212.09611.abs/2212.09611
Hartmann J, Heitmann M, Siebert C, Schamp C (2022) More than a feeling: accuracy and application of sentiment analysis. Int J Res Mark. https://doi.org/10.1016/j.ijresmar.2022.05.005
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778
Hutto C, Gilbert E (2014) May Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media 8(1):216–225. https://doi.org/10.1609/icwsm.v8i1.14550
Jacobs A, Kinder A (2019) Computing the Affective-Aesthetic potential of literary texts. Artif Intell 1:11–27. https://doi.org/10.3390/ai1010002
James D, James A (2020) Symmetry in European regional folk dress: A multidisciplinary analysis. Leonardo 53:157–166
James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer, New York, p 18
Jordan E, Spencer D, Prayag G (2019) Tourism impacts emotions and stress. https://doi.org/10.1016/J.ANNALS.2019.01.011. Annals of Tourism Research
Kaneko H (2023) Interpret machine learning models for data sets with many features using feature importance. ACS Omega 8(25):23218–23225. https://doi.org/10.1021/acsomega.3c03722
Köppen M (2000), September The curse of dimensionality. In: 5th online world conference on soft computing in industrial applications (WSC5) 1:4–8
Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90. https://doi.org/10.1145/3065386
Laing J (2021) Making sense of shame. Philosophy 97:233–255. https://doi.org/10.1017/S0031819121000395
Lever J, Krzywinski M, Altman N (2016) Points of significance: logistic regression. Nat Methods 13:541–542. https://doi.org/10.1038/nmeth.3904
Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW (2019) Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557
Li N, Yang X, Wong I, Law R, Xu J, Zhang B (2022) Automating tourism online reviews: a neural network based aspect-oriented sentiment classification. J Hospitality Tourism Technol. https://doi.org/10.1108/jhtt-03-2021-0099
Li Y, Xu B, Liu Y (2024) A study on the visual comfort of urban Building colors under overcast and rainy weather. https://doi.org/10.3390/buildings14061552. Buildings
Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv Neural Inf Process Syst 32

Marder B, Erz A, Angell R, Plangger K (2019) The role of photograph aesthetics on online review sites: effects of Management- versus Traveler-Generated photos on tourists' decision making. J Travel Res 60:31–46. https://doi.org/10.1177/0047287519895125

Mehra P (2023) Unexpected surprise: emotion analysis and aspect based sentiment analysis (ABSA) of user generated comments to study behavioral intentions of tourists. Tourism Manage Perspect. https://doi.org/10.1016/j.tmp.2022.101063

Meng X, Zhou J, Liu Y, Qiu S (2024) Research on tourists' sentiment tendency to scenic spots based on the transformer deep learning model: A case study of Jade Dragon snow mountain. 2024 IEEE 6th Adv Inform Manage Communicates Electron Autom Control Conf (IMCEC) 6:1248–1253. https://doi.org/10.1109/IMCEC59810.2024.10575891

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of word representations in vector space. ArXiv Preprint arXiv:13013781 https://doi.org/10.48550/arXiv.1301.3781

Nawijn J, Biran A (2019) Negative emotions in tourism: a meaningful analysis. Curr Issues Tourism 22:2386–2398. https://doi.org/10.1080/13683500.2018.1451495

Ng S, Lim K, Lee C, Lim J (2023) Sentiment Analysis using DistilBERT. 2023 IEEE 11th Conference on Systems, Process & Control (ICSPC), 84–89. https://doi.org/10.1109/ICSPC59664.2023.10420272

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M (2021) GLIDE: towards photorealistic image generation and editing with Text-Guided diffusion models., 16784–16804. *ArXiv Preprint arxiv*: https://arxiv.org/abs/2112.10741

Nixon A, Robillard R, Leveille C, Douglass A, Porteous M, Veitch J (2023) Assessing the effects of polychromatic light exposure on mood in adults: A systematic review contrasting α-opic equivalent daylight illuminances. LEUKOS 20:127–147. https://doi.org/10.1080/15502724.2023.2219017

Noira Y, Surayyo M, Muborak N, Shakhnoza S (2021) Lexical borrowing in tourism terminology. 58:309–314. https://doi.org/10.17762/PAE.V58I2.1761

Oliveira T, Araujo B, Tam C (2020) Why do people share their travel experiences on social media? Tour Manag. https://doi.org/10.1016/J.TOURMAN.2019.104041

Park S (2024) Strategies for coping with business travel stressors: enhancing business travel satisfaction through leisure activities. Int J Tourism Res. https://doi.org/10.1002/jtr.2683

Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. 1532–1543. https://doi.org/10.3115/v1/D14-1162

Priya D, Udayan J (2020) Transfer learning techniques for emotion classification on visual features of images in the deep learning network. Int J Speech Technol 23:361–372. https://doi.org/10.1007/s10772-020-09707-w

Pryma V (2023) Tourism lexic in formal and informal discourse. Writings Romance-Germanic Philology. https://doi.org/10.18524/2307-4604.2023.1(50).285563

Rabinowicz A, Rosset S (2022) Tree-based models for correlated data. J Mach Learn Res 23(258):1–31

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sutskever I (2021), July Learning transferable visual models from natural language supervision. In International conference on machine learning, (pp. 8748–8763). PMLR

Rasul T, Santini F, Lim W, Buhalis D, Ramkissoon H, Ladeira W, Pinto D, Azhar M (2024) Tourist engagement: toward an integrated framework using meta-analysis. J Vacation Mark. https://doi.org/10.1177/13567667241238456

Regan C, Iwahashi N, Tanaka S, Oka M (2024) Can generative agents predict emotion? ArXiv. https://doi.org/10.48550/arXiv.2402.04232.abs/2402.04232.

Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-Networks. 3980–3990. https://doi.org/10.18653/v1/D19-1410

Rezaeinia S, Rahmani R, Ghodsi A, Veisi H (2019) Sentiment analysis based on improved pre-trained word embeddings. Expert Syst Appl 117:139–147. https://doi.org/10.1016/j.eswa.2018.08.044

Rico T (2021) Heritage preservation in religious contexts. Disciplinary challenges for the middle East and North Africa (MENA) region. Archaeol Dialogues 28:111–120. https://doi.org/10.1017/S1380203821000143

Ruiz EC, De la Cruz ERR, Vázquez FJC (2019) Sustainable tourism and residents' perception towards the brand: the case of Malaga (Spain). Sustainability 11(1):292. https://doi.org/10.3390/su11010292

Saeed W, Omlin C (2021) Explainable AI (XAI): A systematic Meta-Survey of current challenges and future opportunities. Knowl Based Syst 263:110273. https://doi.org/10.1016/j.knosys.2023.110273

Sanh V (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://arxiv.org/abs/1910.01108v4
Ship E, Spivak E, Agarwal S, Birman R, Hadar O (2024) Real-Time weather image classification with SVM. ArXiv. https://doi.org/10.48550/arXiv.2409.00821.abs/2409.00821
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. ArXiv Preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
Singhal P, Walambe R, Ramanna S, Kotecha K (2023) Domain adaptation: challenges, methods, datasets, and applications. IEEE Access 11:6973–7020. https://doi.org/10.1109/ACCESS.2023.3237025
Song X, Salcianu A, Song Y, Dopson D, Zhou D (2020) Fast WordPiece Tokenization 2089–2103. https://doi.org/10.18653/v1/2021.emnlp-main.160
Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 2818–2826). https://doi.org/10.1109/CVPR.2016.308
Tan M, Le Q (2019), May Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105–6114). PMLR
Valdez P, Mehrabian A (1994) Effects of color on emotions. J Exp Psychol Gen 123(4):394–409. https://doi.org/10.1037/0096-3445.123.4.394
Vargas M, Cannon R, Engel A, Sarwate A, Chiang T (2024) Understanding generative AI content with embedding models. ArXiv. https://doi.org/10.48550/arXiv.2408.10437. abs/2408.10437
Wang Z (2024) AI-based text-to-image synthesis: A review. Appl Comput Eng. https://doi.org/10.54254/2755-2721/45/20241038
Wang J, Li Y, Wu B, Wang Y (2020) Tourism destination image based on tourism user-generated content on internet. Tourism Rev. https://doi.org/10.1108/tr-04-2019-0132
Wang J, Liu H, Jing L (2024) Transparent embedding space for interpretable image recognition. IEEE Trans Circuits Syst Video Technol 34:3204–3219. https://doi.org/10.1109/TCSVT.2023.3314769
Warstadt A, Zhang Y, Li H-S, Liu H, Samuel R (2020) Bowman. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). arXiv preprint arXiv:2010.05358
Wattanacharoensil W, La-Ornual D (2019) A systematic review of cognitive biases in tourist decisions. Tour Manag 75:353–369. https://doi.org/10.1016/J.TOURMAN.2019.06.006
Wen T, Xu X (2024) Research on image perception of tourist destinations based on the BERT-BiLSTM-CNN-Attention model. Sustainability. https://doi.org/10.3390/su16083464
Wu L, Wu C, Guo H, Zhao Z (2023) A Cross-Modal alignment for Zero-Shot image classification. IEEE Access 11:9067–9073. https://doi.org/10.1109/ACCESS.2023.3237966
Wu T, Tao C, Wang J, Zhao Z, Wong N (2024) Rethinking Kullback-Leibler divergence in knowledge distillation for large Language models. ArXiv abs/240402657. https://doi.org/10.48550/arXiv.2404.02657
Xia X, Xu C, Nan B (2017) Inception-v3 for flower classification. 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 783–787. https://doi.org/10.1109/ICIVC.2017.7984661
Xu J, Du Q (2019) A deep investigation into fasttext. 2019 IEEE 21st Int Conf High-Performance Comput Communications; IEEE 17th Int Conf Smart City; IEEE 5th Int Conf Data Sci Syst (HPCC/SmartCity/DSS) 1714–1719. https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00234
Yin C, Zhang S, Zeng Q (2023) Hybrid representation and decision fusion towards Visual-textual sentiment. ACM Trans Intell Syst Technol 14:1–17. https://doi.org/10.1145/3583076
Yu N, Lv Y, Liu X, Jiang S, Xie H, Zhang X, Xu K (2023) Impact of correlated color temperature on visitors' perception and preference in virtual reality museum exhibitions. Int J Environ Res Public Health 20. https://doi.org/10.3390/ijerph20042811
Zaino G, Recchiuto C, Sgorbissa A (2022) Culture-to-Culture image translation with generative adversarial networks. ArXiv, abs/2201.01565. https://doi.org/10.48550/arXiv.2201.01565
Zhang H, Dong J, He S, Lv S (2020) Research on Image Complexity Description Method Based on Approximate Entropy. 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 918–920. https://doi.org/10.1109/ICITBS49701.2020.00203
Zhang Z, Zhuo K, Wei W, Li F, Yin J, Xu L (2021) Emotional responses to the visual patterns of urban streets: evidence from physiological and subjective indicators. Int J Environ Res Public Health 18. https://doi.org/10.3390/ijerph18189677
Zhao Z, Zhu H, Xue Z, Liu Z, Tian J, Chua M, Liu M (2019) An image-text consistency driven multimodal sentiment analysis approach for social media. Inf Process Manag 56. https://doi.org/10.1016/J.IPM.2019.102097

Authors and Affiliations

Víctor Calderón-Fajardo[1,2] · Ignacio Rodríguez-Rodríguez[1] · Miguel Puig-Cabrera[3]

✉ Víctor Calderón-Fajardo
vcalde02@ucm.es

✉ Ignacio Rodríguez-Rodríguez
ignacio.rodriguez@ic.uma.es

✉ Miguel Puig-Cabrera
mpcabrera@ualg.pt

[1] University of Malaga, Málaga, Spain

[2] Complutense University of Madrid, Madrid, Spain

[3] University of Algarve, Faro, Portugal