

Article

A Transformer-Based Semantic Encoding Framework for Quantitative Analysis of Large-Scale Textual Reviews

Darjan Karabašević ^{1,2,*} , Aleksandra Vujko ^{3,*} , Vuk Mirčetić ² , Gabrijela Popović ^{2,4} 
and Dragiša Stanujkić ^{5,6} 

¹ College of Global Business, Korea University, Sejong 30019, Republic of Korea

² Faculty of Applied Management, Economics and Finance in Belgrade, University Business Academy in Novi Sad, Jevrejska 24, 11000 Belgrade, Serbia

³ Faculty of Tourism and Hospitality Management, Singidunum University, 11000 Belgrade, Serbia

⁴ Department of Mathematics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai 602105, India

⁵ University College, Korea University, 145 Anam-ro, Seoul 02841, Republic of Korea

⁶ Technical Faculty in Bor, University of Belgrade Vojske, Jugoslavije 12, 19210 Bor, Serbia

* Correspondence: darjankarabasevic@korea.ac.kr (D.K.); avujko@singidunum.ac.rs (A.V.)

Abstract

Increasing turbulence in contemporary business environments has made the quantitative analysis of unstructured textual data a central methodological challenge for researchers and decision-makers. The increasing availability of large-scale textual data has heightened the need for quantitative frameworks that can transform unstructured language into analyzable numerical representations. Transformer-based language models address this need by encoding text into high-dimensional semantic embeddings. Yet, these representations are commonly treated as black-box inputs for downstream tasks, with limited examination of their intrinsic numerical and geometric properties. The research in this manuscript addresses this gap by proposing a quantitative framework for analyzing transformer-based semantic embeddings as high-dimensional metric spaces prior to task-specific modeling. We employ an innovative methodological approach, considering vector norms regarding examining the dispersion of vector norms to detect concentration of measure, cosine similarity in the context of evaluating the distribution of pairwise cosines between vectors, and principal component analysis. For the purpose of the research, 3034 visitor-generated reviews related to national park experiences were used. Textual inputs are deterministically mapped into a normalized 384-dimensional embedding space using a transformer-based encoder. The analysis examines numerical stability through vector norm dispersion, semantic organization via cosine similarity distributions, variance structure using principal component analysis, and internal organization through unsupervised clustering validity metrics. Clustering is successful when high separation between clusters and high cohesion within clusters are achieved, which is why a single measure combining separation and cohesion metrics was proposed in the research. The results show almost perfect norm stability, backing up the choice of angular similarity as the right semantic metric. Variance decomposition and clustering results share a continuous high-dimensional semantic structure with no dominant latent components or clearly separable clusters. These results suggest that semantic meaning is best thought of as a continuous metric space rather than discrete categories, highlighting the need for representational diagnostics before predictive modeling.

Keywords: transformer-based embeddings; semantic encoding; high-dimensional vector spaces; cosine similarity; principal component analysis; clustering validity; quantitative text analysis; metric space analysis



Academic Editor: Stefano De Marchi

Received: 13 January 2026

Revised: 14 February 2026

Accepted: 26 February 2026

Published: 28 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

MSC: 68T50; 62H25

1. Introduction

The proliferation of user-generated text on digital platforms has increased both the volume and analytical relevance of unstructured data [1–3]. Despite the abundance of such texts, transformer-based embeddings—though widely adopted—are often treated as black-box features, and their internal numerical and geometric properties remain largely unexplored [4]. This creates a methodological gap that this study aims to address. While these texts are rich in semantic and experiential content, their unstructured and heterogeneous nature makes conventional quantitative analysis challenging [5–7]. In domains like tourism and protected area management, visitor narratives often integrate evaluative judgments and situational descriptions, making simple text-to-number transformations insufficient. Traditional text representation methods, such as bag-of-words, TF-IDF, and probabilistic topic models, rely primarily on lexical features and are insensitive to word order or contextual nuance [8–10]. This limits their ability to capture the complexity of natural language, especially in domains like tourism, where visitor narratives combine affective reactions, evaluative judgments, and situational descriptions. Reducing such texts to isolated word counts risks losing analytically important dimensions of meaning.

Recent advances in transformer-based language models have shifted text representation from isolated tokens to dense, context-sensitive embeddings [11,12]. These embeddings perform reliably across downstream tasks and have been widely adopted in tourism research for sentiment analysis and decision-support applications [13,14]. However, their rapid uptake has created a methodological blind spot: embeddings are often treated as black-box features, and their internal numerical and geometric structure remains largely unexplored. Embeddings are often treated as black-box features, serving as inputs to predictive or classification models without examining their internal structure [15]. As a result, key properties such as numerical consistency, geometric organization, and variance behavior remain largely unexplored [16]. This lack of understanding limits the reliability of embedding-based analyses in applied settings, including national parks, where misinterpretation of semantic representations can affect experience design, crowd management, and decision-making [17].

Many statistical and machine-learning methods implicitly assume well-behaved metrics, variance distributions, and regular data structures [18]. Introducing semantic embeddings without first examining their numerical and geometric properties leaves these assumptions untested, especially in high-dimensional contexts where variability concentration, distance concentration, and weak cluster separability can distort similarity measures and downstream analyses [19]. Existing evaluations often focus on predictive performance, implicitly equating accuracy with representational adequacy [20–24]. While suitable for task-specific objectives, this approach overlooks whether embeddings reliably capture the internal organization of semantic space. In national park reviews, visitor experiences are complex, involving mixed perceptions and trade-offs that simple task-based evaluation may fail to reflect.

The research conducted in this manuscript addresses the identified gap by developing a measurement structure for investigating transformer-based semantic embeddings as high-dimensional metric spaces [25]. Visitor reviews from national parks are used because they represent a wide range of heterogeneous expressions. The objective of this study is not empirical generalization across tourism contexts, but methodological generalization, demonstrating how semantic embedding structures can be examined independently of

domain-specific content. The manuscript treats semantic embeddings as numerical representations whose internal properties deserve attention in their own right, rather than focusing on task-specific prediction or domain-driven interpretation. The analysis concentrates on established statistical and geometric descriptors—including vector norms, cosine similarity distributions, variance structures explored through dimension reduction, and clustering validity metrics—examined prior to any predictive or optimization-oriented modeling. Taken together, this framework enables explicit assessment of numerical stability, geometric coherence, and semantic organization, providing insights into how visitor experiences are structured within the embedding space.

This manuscript contributes to the quantitative analysis of unstructured textual data by proposing a novel analytical framework for examining semantic embeddings. Unlike conventional embedding-based studies that rely on clustering, topic extraction, or similarity ranking, this approach treats embeddings as continuous high-dimensional metric spaces. It examines relational distance structures and dispersion patterns, interpreting these properties as indicators of latent experiential organization rather than thematic prevalence. Conceptually, semantic encoding is treated as a deterministic mapping from unstructured visitor narratives to structured numerical space, enabling quantitative examination of experiential meaning. The study assesses the numerical and geometric properties of transformer-based embeddings applied to heterogeneous national park reviews, focusing on internal organization and stability. This approach emphasizes similarity relations, variance behavior, and organizational coherence rather than assuming adequacy based on predictive performance.

Building on these results, the study proposes a methodological framework that can be applied prior to task-based analytics, facilitating more informed interpretation of sentiment classification, segmentation, and decision-support models in protected area management. Overall, this strengthens the methodological basis for integrating unstructured textual data while retaining the semantic nuances in visitor narratives. The manuscript is organized as follows: the next chapter provides research background and related work; the methodological framework is then presented, including data description, vector norms, cosine similarity, principal component analysis, and clustering validity metrics. The following chapter presents empirical results, followed by a discussion of methodological and managerial implications. The final chapter concludes with limitations and directions for future research.

2. Background and Related Work

Recent progress in language processing is based on models that can understand the meaning of text depending on context, using large amounts of data [26]. Earlier approaches, such as simple word-count models or fixed word vectors, could not capture meaning in this way [27,28]. Modern models can link distant parts of a text and interpret the meaning of the whole message, not just individual words [28,29]. One practical result of this development is that text can be converted into numbers and represented in a shared “meaning space.” This makes it possible to analyze meaning quantitatively, not just by comparing individual words [30]. Sentence-level models go a step further by representing an entire sentence or short text with a single set of numbers. This allows large amounts of user-generated text to be analyzed while still preserving the overall meaning, even when expressions vary widely [31,32].

These models are now used in many areas, including text meaning comparison, grouping similar content, information retrieval, and general text analysis [33]. Their use is particularly visible in tourism and recreation research, where they are increasingly applied to analyze visitor reviews, satisfaction, and descriptions of experiences in tourist

destinations and protected areas [34]. National parks have proven especially well-suited to such approaches. The volume and richness of visitor-generated textual data capture diverse perceptions of nature quality, infrastructure, accessibility, crowding, interpretation, and overall experience [35]. Most studies evaluate these models based on how well they perform specific tasks, such as identifying whether someone is satisfied or dissatisfied, or supporting decision-making [36]. While this is useful in practice, it tends to overlook how meanings are actually structured within the models themselves.

Similarity in meaning between texts can be reliably measured using cosine similarity, a measure that indicates how similar two texts are in meaning regardless of their length, as well as related methods [37]. This is particularly useful when working with large and complex datasets [38,39]. In applied research, these measures are often used to identify similar reviews, aggregate visitor narratives, and support recommendation and monitoring systems [40]. Their usefulness in such contexts is well established. At the same time, many of these task-oriented applications rely on assumptions that are rarely examined explicitly. Numerical stability, spatial coherence, and the overall structure of the embedding space are usually taken for granted rather than carefully tested [41,42]. Embedding representations are often used as a reliable basis for analysis. However, it is rarely examined how different meanings are actually arranged within that space, or how similar, different, or well-organized they are.

Techniques for dimension reduction, including principal component analysis (PCA), are typically used for initial visualization of the embedding space and to provide a first intuition for how variance is concentrated [43,44]. In practical/applied research, people often interpret low-dimensional projections as major experiential themes or latent factors underlying visitor discourse [45,46]. Such an interpretation appears intuitive and simple when multifaceted datasets are reduced to simpler forms that can be scrutinized and communicated. However, in fact/practice, the application is largely descriptive. Formal discussion of how semantic information is distributed across latent dimensions, or of how much variance must be retained to preserve meaningful structure, is often limited or absent [47]. That means structure is only partially or contextually defined by low-dimensional projections. Practically, in the case of national park studies, this result has consequences. Most often, visitor accounts are complex and ambiguous; positivity and negativity can easily be merged within one single account. Therefore, when such experiential narratives are reduced to a few components for purposes of visualization or reporting, the apparent experiential patterning will almost certainly be an oversimplification of the gradual, overlapping nature perceptions of visitors.

Unsupervised clustering methods are also widely applied to semantic embeddings to extract themes or visitor segments [48]. The typical interpretation of such clusters, in tourism research and park management studies, is that they represent distinct types of visitors or experiences [49]. This idea is appealing because it suggests that clearly defined groups can be directly translated into practical managerial decisions. In reality, however, little practical work has been conducted on checking the validity of clustering against alternative configurations. Very often, for reasons of interpretative convenience and not because it is strictly necessary from a methodological point of view, weak overlapping or continuous structures are forced into discrete partitions [50,51]. There is a risk in this practice. It may exaggerate categorical distinctions in visitor experiences that are, in reality, organized along slow semantic gradients by overlapping motivations, perceptions, and context-dependent trade-offs. In the simplest terms, this shows that there is a general problem. Although transformer-based semantic representations are now widely used and increasingly applied in decision-support systems [52,53], there is still no clear and established way to analyze them quantitatively. Embedding spaces are mostly treated as a

technical step in the analysis, and their value is judged by how well they perform a given task rather than by examining their numerical and spatial structure [54]. As a result, some important patterns in these representations may go unnoticed, even though their outputs are used in real-world decision-making.

This limitation becomes particularly significant in practical applications, such as national park management, where insights derived from textual data analysis support decision-making at multiple levels, with an emphasis on sustainable and caring capacity management [55]. Good performance on standard measures does not necessarily mean that all crucial aspects or dimensions of visitors' experience have been captured. Temporary experiences, mixed opinions, or early signs of problems often do not fit into predefined categories and can therefore be easily overlooked [56,57]. Such blind spots are hard to identify without analyzing the embedding space's actual structure. High-performing models mask subtle but meaningful variation in visitor discourse and appear reliable. This observation speaks to a higher-level methodological need that is more broadly applicable. It is desirable to apply approaches before building specific models that carefully examine the numerical stability of semantic representations, as well as their structure and grouping, rather than assuming they function correctly on their own.

This study responds to that methodological need by proposing a clearer, more systematic quantitative approach to analyzing transformer-based semantic embeddings, treating them as complex, high-dimensional spaces. Rather than focusing on how successful a model is at a specific task or domain, the analysis examines the internal structure of semantic embeddings and how it influences subsequent analyses. By relying on clear and measurable analytical procedures, the study seeks to clarify how visitor experiences are encoded and organized within complex, high-dimensional semantic embeddings. This approach does not replace task-based analyses; instead, it complements them. It introduces a level of structural validation that enables more cautious and transparent interpretation of forecasting and classification results. In applied contexts, such as national park management, and in related domains where AI-generated textual analytics inform decision-making. The mentioned shift supports a more grounded interpretation and application of model outputs.

To contextualize this methodological position, Table 1 provides a structured overview of prior studies that use semantic embeddings to analyze unstructured textual data. The table compares these studies by their research objectives, embedding models, analytical techniques, and key findings. This comparison enables identification of common methodological patterns and recurring limitations in existing approaches.

Table 1. Overview of prior studies using semantic embeddings for text analysis.

Study (Reference)	Research Objective	Embedding/Representation Model	Analytical Methods	Key Findings	Identified Limitation
[1]	Review of social media analytics applications	Various text representations	Descriptive review	Demonstrates the growing importance of UGC analytics	Focus on applications, not embedding structure
[2]	Systematic review of UGC text mining in e-commerce	TF-IDF, word embeddings	Topic modeling, classification	Identifies dominant application areas	Embeddings treated as task inputs
[3]	Topic modeling in tourism research	Probabilistic topic models	Topic extraction	Managerial insights from themes	Ignores semantic geometry
[5]	Structuring unstructured feedback using LLMs	Transformer-based embeddings	Task-oriented modeling	Improved value proposition modeling	No analysis of the embedding space

Table 1. *Cont.*

Study (Reference)	Research Objective	Embedding/Representation Model	Analytical Methods	Key Findings	Identified Limitation
[9]	ML prediction from idiographic text	Static and contextual embeddings	Prediction accuracy evaluation	Shows limits of ML-based inference	No spatial analysis of representations
[13]	Emotion and perception analysis in parks	Deep contextual embeddings	Sentiment analysis	Reveals emotional patterns	Relies on downstream performance
[34]	Topic mining in forest recreation research	Contextual embeddings	Topic modeling	Semantic-driven topic extraction	Structural properties unexplored
[45]	Hotel robot perception analysis	Transformer-based embeddings	PCA + sentiment classification	Identifies satisfaction dimensions	PCA was used descriptively only
[52]	Embedding-based decision support	Contextual embeddings	Clustering and prediction	Supports large-scale decisions	No validation of space geometry
[17]	Interpretation of embedding dimensions	Language model embeddings	Knowledge-graph alignment	Improves interpretability	Focused on mapping, not geometry

As summarized in Table 1, existing studies predominantly evaluate semantic embeddings through downstream task performance or descriptive projections. Meanwhile, systematic examination of numerical stability, geometric organization, and variance structure remains largely absent. This gap motivates the analytical framework introduced in the following section, which treats semantic embeddings as objects of quantitative investigation in their own right.

3. Materials and Methods

3.1. Data Source and Corpus Description

The dataset was selected for its semantic richness and heterogeneity, making it an ideal testbed for examining transformer-based embeddings. Using national park reviews allows us to explore how experiential meaning is encoded in high-dimensional semantic spaces, independently of domain-specific content, demonstrating the broader applicability of the proposed analytical framework. The dataset was obtained from an open-access online repository, specifically the Kaggle platform, where it is distributed under the title Parks Reviews Dataset. The dataset used by the authors in this study comprises visitor reviews of national parks published during 2025. It serves as the empirical basis for the research, in which textual representations and transformer-based semantic modeling are examined. National park reviews are used as a theoretically suitable testbed due to their semantic richness; however, the proposed analytical framework is not domain-bound and can be applied to any corpus of unstructured experiential text. The dataset consists of 3034 English-language reviews that vary considerably in length, structure, and expressive complexity. Each review is treated as an independent textual observation and contains unstructured natural-language content reflecting visitors' subjective evaluations, perceptions, and experiential accounts of park environments. In practice, these accounts rarely focus on a single aspect of the visit. Rather than focusing solely on the natural features of national parks and their undeniable beauty, the reviews typically combine descriptions of the landscape with evaluations of infrastructure, accessibility, crowding, the quality of visitor information, safety, and overall satisfaction. Taken as a whole, these texts provide a clear and balanced overview of visitor experiences in national parks. In this context, meaning does not emerge from isolated elements, but from the interaction between environmental conditions, park organization, and visitors' emotional responses. Precisely because of this diversity, the dataset is particularly well suited for examining how meaning is encoded and organized within high-dimensional semantic spaces, without imposing predefined categories or interpretive constraints.

From a methodological perspective, national park reviews constitute a particularly demanding form of textual data. The way visitors express their views on protected areas shows considerable diversity in motivations, levels of environmental awareness, and styles of expression. Some reviews take the form of brief evaluations, while others are more extensive and written as narrative accounts. Visitors to national parks often include both positive and negative evaluations within the same review, while reviews that are entirely positive or entirely negative are very rare. Because of this diversity, the dataset is well-suited for analysis using semantic models that examine how meaning is constructed and connected into a coherent whole. Encoding models must be able to process all types of responses within a single numerical framework. In this study, such diversity is not treated as noise that should be removed or reduced. Instead, it is regarded as a central feature that allows latent semantic structure to emerge organically from the data.

Before analysis, the dataset was reviewed, and only records containing actual textual content were retained; entries without text and partial records lacking meaningful content were removed. Importantly, no additional filtering based on textual meaning or content was applied. By avoiding predefined categories during data preparation, the full diversity of how visitors describe their experiences is preserved. Rather than adapting the texts to a specific analytical objective, the dataset is treated as a direct, unfiltered record of visitors' experiences in national parks. This approach allows meaning to emerge directly from the text and enables examination of how it is organized within semantic models, without additional assumptions or constraints.

3.2. Methodology

The overall analytical workflow is summarized in Figure 1, illustrating the sequence from text collection and embedding generation to geometric and statistical diagnostics.

Figure 1 illustrates the methodological sequence from text collection and transformer-based embedding generation to vector normalization and exploratory geometric diagnostics, including cosine similarity analysis, principal component analysis, and clustering validity assessment. These components jointly support the interpretation of semantic space structure prior to task-specific modeling. The mathematical representation of the applied methodological approach for analyzing unstructured textual data is presented in Section 3.2.

3.2.1. Vector Norms

A vector norm denotes a function $f : V \rightarrow \mathbb{R}$ over a vector space V , such that for $x, y \in V$ and a scalar $a \in \mathbb{R}$, the following characteristics are established [58]:

$$f(ax) = |a|f(x), \quad (1)$$

$$f(x + y) \leq f(x) + f(y), \quad (2)$$

$$f(x) \geq 0, \quad (3)$$

if $f(x) = 0$, then $\mathbf{0} = \mathbf{0}$, which is the zero vector.

The l_p vector norm is defined in the following way:

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad p \geq 1. \quad (4)$$

The taxicab (Manhattan grid) exists when $p = 1$:

$$\|x\|_1 = \sum_{i=1}^n |x_i|. \quad (5)$$

Euclidean norm is in the case when $p = 2$:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}. \quad (6)$$

The maximum norm for $p = \infty$ is as follows:

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_n|\}. \quad (7)$$

Norm-based dispersion that outlines a relative spread of components with respect to the vector magnitude could be as follows:

$$d(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{x_i}{\|x\|} - \frac{1}{n} \sum_{j=1}^n \frac{x_j}{\|x\|} \right)^2}. \quad (8)$$

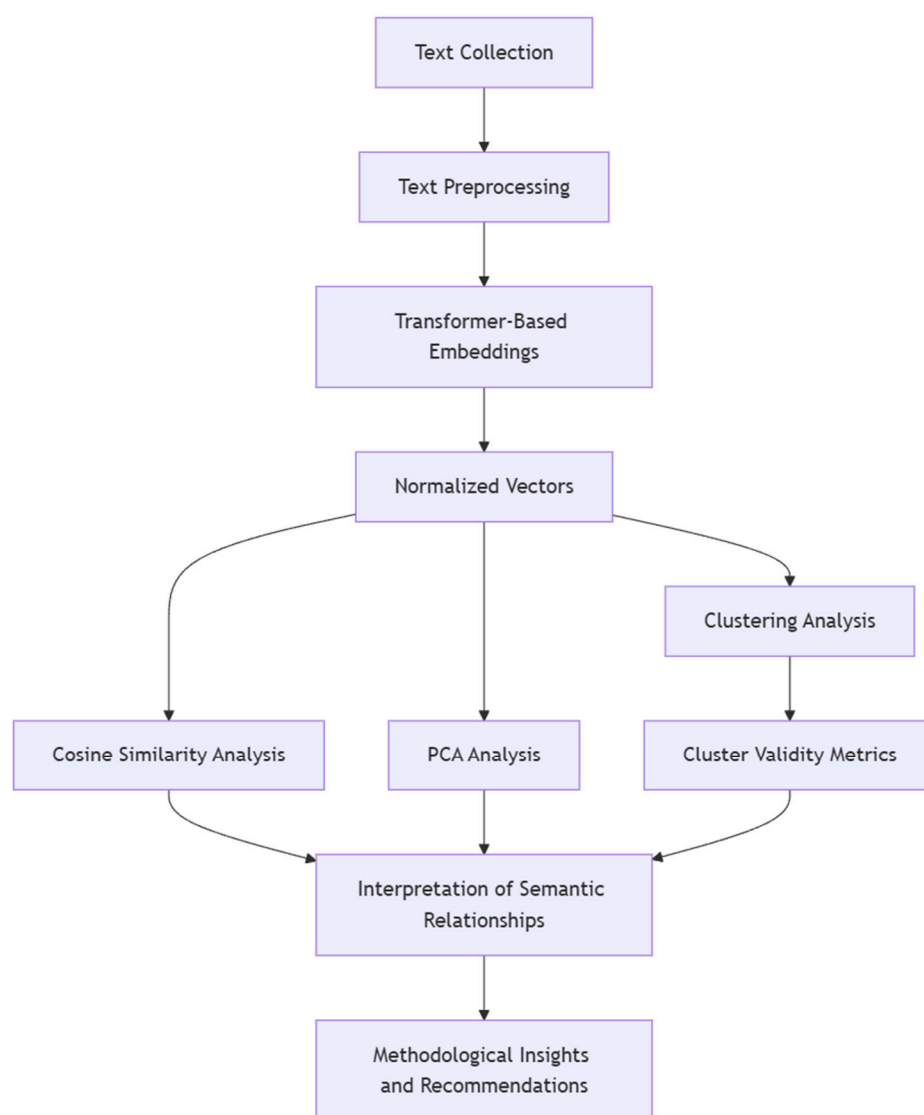


Figure 1. Analytical workflow of the proposed framework.

3.2.2. Cosine Similarity

The angle between two vectors is estimated using the cosine similarity. This measure is calculated by dividing the vectors' dot product by the product of their norm and could be computed as follows:

$$\text{cosim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}. \quad (9)$$

The range of cosine similarity is between 0 and 1. When the cosine similarity is close to 1, the two vectors are considered similar. In the opposite case, when cosine similarity is close to 0, two vectors are considered dissimilar [59].

3.2.3. Principal Component Analysis

Principal Component Analysis (PCA) reduces the data to a few specific dimensions by computing a new set of variables. Each new variable represents a linear combination of the previous variables. A more reliable approximation is determined using more principal components, where every succeeding component is uncorrelated with the previous one. Therefore, the formula for k principal components is written as follows:

$$X = PA^T + \epsilon, \quad (10)$$

where X denotes the $m \times n$ matrix of original variables. P represents $n \times k$ matrix of new principal components, where $k \leq m$. A^T denotes $k \times n$ matrix of coefficients in relation to the original variables to the k principal components. ϵ represents the residual matrix (with $\epsilon = 0$ if $k = n$) [60].

3.2.4. Unsupervised Clustering Validity Metrics

The clustering target is very precise—we aimed to perform an adequate clustering of the unlabeled dataset. Clustering is successful when high separation between clusters and high cohesion within clusters are achieved. The authors proposed a single measure combining separation and cohesion metrics. The most famous measures are shown below [61,62].

The Calinski–Harabasz coefficient (variance ratio criterion) represents a measure based on the internal cluster’s dispersion and the dispersion between clusters. Mathematically, it can be written as follows:

$$CH = \frac{\frac{SSB_M}{M-1}}{\frac{SSE_M}{(M)}}. \quad (11)$$

The Dunn index represents a ratio of the smallest and the largest distance between data from different clusters, and it is represented in the following way:

$$D = \min_{1 \leq i < k} \left\{ \min_{1 \leq j < k, i \neq j} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \leq l < k} \{\Delta(C_l)\}} \right\} \right\}, \quad (12)$$

$$\Delta(C_i) = \max_{x, y \in C_i} \{d(x, y)\}, \quad (13)$$

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{d(x, y)\}. \quad (14)$$

The Xie–Beni score was designed for fuzzy clustering, but it applies to crisp clustering, and it could be presented as follows:

$$XB = \frac{\sum_{i=1}^N \sum_{k=1}^M u_{ik}^2 \|x_i - C_k\|^2}{N_{t \neq s} \min \{ \|C_t - C_s\|^2 \}}. \quad (15)$$

The Ball–Hall index represents a dispersion measure based on the quadratic distances of the cluster points from the centroid:

$$BH = \frac{SSE_M}{M}. \quad (16)$$

Another measure is a Hartigan index that is based on the logarithmic relationship between the sum of squares within and between the clusters:

$$H = \log\left(\frac{SSB_M}{SSE_M}\right). \quad (17)$$

The silhouette coefficient is the most popular method for estimating separation and cohesion metrics, and it is calculated in three steps.

Step 1. Compute the average distance $a(i)$ for all points in the same cluster for all clusters:

$$a(i) = \frac{1}{|C_a|} \sum_{j \in C_a, i \neq j} d(i, j). \quad (18)$$

Step 2. Compute the minimum average distance $b(a)$ between the considered point and the points that are in each cluster, omitting the analyzed point:

$$b(i) = \min_{C_b \neq C_a} \frac{1}{|C_b|} \sum_{j \in C_b} d(i, j). \quad (19)$$

Step 3. The silhouette coefficient is determined by using the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (20)$$

The global silhouette coefficient is calculated as the average of the silhouette coefficients for each point in the dataset:

$$S = \frac{1}{n} \sum_{i=1}^n s(i). \quad (21)$$

3.3. Text Preprocessing

Text preprocessing prior to analysis was kept to a minimum in order to allow the underlying structure of transformer-based semantic embeddings to be examined more clearly. Visitor reviews were retained in their original form to preserve their meaning, writing style, and narrative flow. This approach treats reviews as holistic descriptions of experience rather than as collections of isolated elements. In this study, diversity in the content and length of reviews is not treated as noise, but as a key characteristic of the data. Within a single review, visitors often combine descriptions of the natural environment, personal impressions, emotional responses, and practical details, with positive and negative experiences frequently appearing together.

By preserving the original form of the text, meaning can emerge from the full context, shaped by the relationships between words and their use within the narrative. Rather than steering the data toward a predefined analytical objective, the analysis focuses on observing how semantic structure emerges when visitor comments are encoded without prior categorization or labeling. Data preparation was therefore limited to technical steps necessary for reliable analysis, including the removal of non-textual metadata and header fields, verification of record integrity, and elimination of empty or malformed textual entries. These procedures do not alter the semantic content of the text, but ensure that all records constitute valid inputs for the semantic encoding process. Preserving differences in review length and writing style enables semantic models to faithfully represent the diversity of visitor experiences in national parks. This is particularly important in practice, as it ensures that insights derived from the analysis reflect real visitor experiences rather than artificially shaped representations.

3.4. Transformer-Based Semantic Encoding

Each textual review was encoded using a pre-trained transformer-based sentence embedding model. The encoder maps each input text $\mathcal{T}_i$ to a fixed-length numerical

vector $e_i \in \mathbb{R}^{384}$, producing a continuous semantic representation suitable for quantitative analysis. Formally, the encoding process can be expressed as follows:

$$f : \mathcal{T} \rightarrow \mathbb{R}^{384} \quad (22)$$

where $\mathcal{T}$ denotes the set of textual inputs and $f(\cdot)$ represents the transformer-based encoding function.

This approach shows how unstructured visitor review texts are translated into a numerical space where meaning can be examined through relationships between points. Reviews are treated as input texts, while their meaning is represented as positions within a shared, continuous space. In this way, meaning, which is usually interpreted in descriptive or qualitative terms, can be analyzed using statistical and spatial methods. At the level of sentences or short texts, the models assimilate word meaning, sentence composition, and wider context. Rather than assigning a single, fixed meaning to each word, transformer-based models interpret meaning in relation to surrounding words. Meaning, therefore, emerges from context, not from isolated terms. This is particularly important for national park reviews. The same words—such as “crowded”, “wild”, “quiet”, or “accessible”—can carry different meanings depending on how they are used. Their interpretation depends on the overall narrative, accompanying words, and the specific circumstance described.

By preserving contextual information, transformer-based models retain these differences in meaning within the semantic space, rather than losing them through simplification or surface-level encoding. To ensure consistent comparison, all vectors were normalized to the same length. This removes technical differences related to text length or writing style, allowing meaning to be compared solely on the basis of content similarity. That is why cosine similarity is used for comparison, as it reflects similarity in the meaning of the texts rather than differences in their length or writing style. This normalization is not a technical formality, but a deliberate methodological choice that allows reviews to be compared based on genuine similarity of experience, rather than on how they are written. The resulting embedding space can therefore be interpreted as a semantic space in which proximity reflects experiential similarity, without reliance on predefined categories or labels.

3.5. Quantitative Analysis of the Embedding Space

The analysis relied on normalized vector representations generated by a transformer-based encoder. Before any further processing, vector norms were checked to make sure that all embeddings were numerically comparable and that no residual magnitude effects remained. Cosine similarity was then used to examine relationships between review embeddings. Rather than inspecting individual text pairs, the focus was placed on the overall distribution of similarity values across the space. This provided a general view of how semantic proximity is structured in the corpus. To inspect how variance is spread across dimensions, principal component analysis (PCA) was applied to the normalized embeddings. In parallel, k-means clustering was used to explore several values of k. No assumptions were made about the existence of predefined semantic groups, and cluster validity was evaluated using standard internal metrics.

4. Results

4.1. Descriptive Characteristics of the Textual Dataset

Table 2 presents the descriptive statistics of the textual dataset and shows that 3034 reviews were analyzed, providing a solid basis for quantitative analysis. The average review length is approximately 48 words, but the data indicate clear variation in how visitors report their experiences. The corpus includes very short comments consisting of only a few words, as well as reviews exceeding 500 words, indicating different writing

styles and levels of detail in visit descriptions. All reviews are written in English; hence, the observed differences can be primarily attributed to content and personal expression. This wide variation in review length and narrative style highlights the heterogeneous nature of visitor-generated text and reinforces the need for representation methods that can capture meaning beyond surface-level lexical patterns. From a methodological perspective, such diversity provides a suitable empirical basis for examining how semantic embeddings accommodate differences in expression while preserving relational coherence in high-dimensional space.

Table 2. Descriptive Statistics of the Textual Dataset.

Measure	Value
Number of textual reviews (N)	3034
Mean review length (words)	47.61
Standard deviation	38.43
Minimum length (words)	4
Maximum length (words)	510
Language	English

4.2. Numerical Stability of the Semantic Embedding Space

This section adopts an exploratory diagnostic perspective; the aim is not hypothesis testing but the interpretation of semantic space configurations revealed by the embedding analysis. Given that embeddings encode high-dimensional relational structures, formal significance testing would be methodologically inappropriate at this stage and would obscure the spatial patterns that constitute the primary analytical contribution. Table 3 reports the numerical properties of the normalized semantic embedding space. Each review is represented by a 384-dimensional vector, which defines the resolution of the semantic representation used in the analysis. The mean vector norm is exactly 1.000, with an extremely small standard deviation, indicating that normalization was applied consistently across the entire corpus. The minimum and maximum norm values differ by only the level of numerical precision, confirming that there is virtually no variation in vector length. This stability shows that differences between embeddings are not driven by magnitude effects, such as text length or verbosity, but solely by their semantic orientation. As a result, comparisons between reviews are based on meaning rather than technical artifacts, providing a reliable numerical foundation for subsequent similarity, variance, and clustering analyses. This result confirms that subsequent analyses reflect structural properties of semantic orientation rather than artifacts introduced by vector magnitude or text length, which is essential for valid geometric interpretation of the embedding space.

All semantic embeddings were L2-normalized prior to analysis. The reported values summarize vector norm statistics across the corpus.

Table 3. Numerical Properties of the Normalized Semantic Embedding Space.

Property	Value
Embedding dimensionality	384
Mean vector norm	1.000
Standard deviation of norms	5.29×10^{-8}
Minimum norm	0.9999998
Maximum norm	1.0000001

4.3. Global Distribution of Semantic Similarity

Table 4 shows how reviews are related to one another in terms of meaning, based on cosine similarity between their semantic representations. The average similarity value is 0.464, indicating that reviews share a common semantic ground. This suggests that visitors often discuss similar aspects of their experiences but describe them from different perspectives. At the same time, there is considerable variation in similarity values. Some reviews are almost identical in meaning, while others show very low or even negative similarity, pointing to clearly different or contrasting descriptions of experiences. This range of values further confirms that visitor discourse covers a wide spectrum of experiences, from very similar to highly divergent. The moderate mean similarity value suggests that visitor reviews share a common experiential vocabulary, while still allowing for substantial semantic differentiation. High similarity values indicate recurring experiential patterns, whereas low or negative similarities reflect contrasting perceptions, expectations, or situational conditions. This distribution supports the view that visitor discourse is neither random nor uniformly homogeneous, but structured around overlapping semantic themes with gradual transitions rather than sharp boundaries.

Table 4. Distribution of Pairwise Cosine Similarities in the Semantic Embedding Space.

Statistic	Value
Mean cosine similarity	0.464
Standard deviation	0.127
Minimum similarity	−0.062
Maximum similarity	1.000
Number of sampled pairs	300,000

Cosine similarity values were computed on randomly sampled pairs of normalized semantic embeddings.

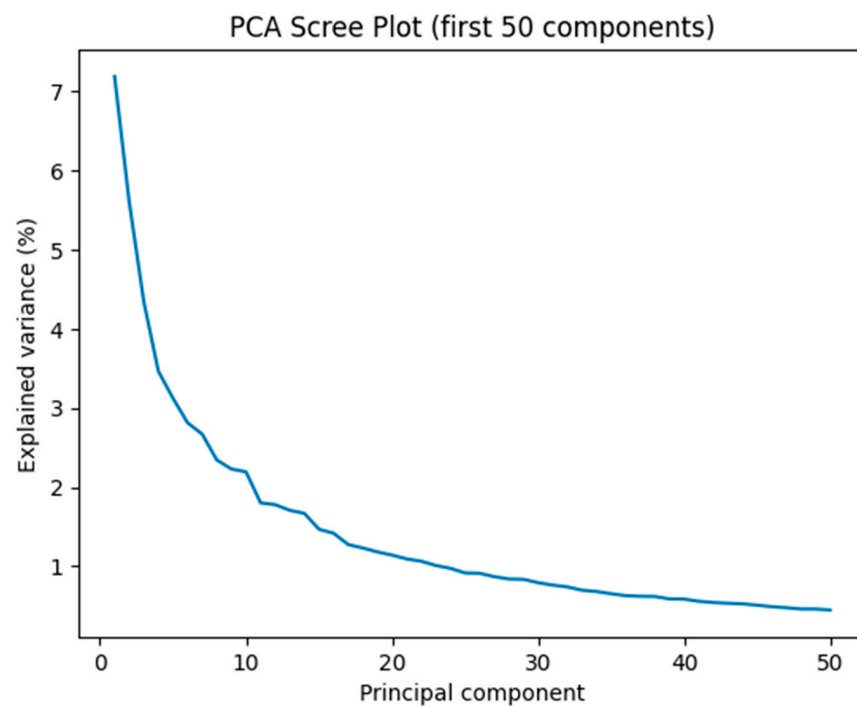
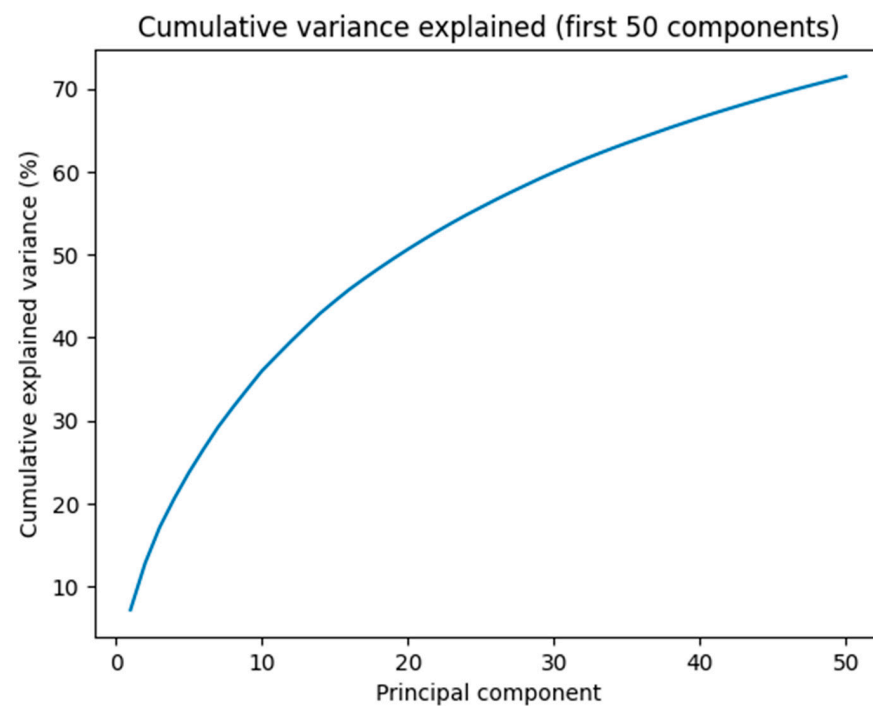
4.4. Latent Dimensional Structure of the Semantic Space

Principal component analysis was applied to the normalized semantic embedding matrix to examine variance distribution across latent dimensions. Table 5 reports explained variance ratios for the first five principal components. The first component accounts for 7.19% of total variance, while the cumulative variance explained by the first five components is 23.73%. As shown in the PCA scree plot (Figure 2) and the cumulative variance curve (Figure 3), variance decays gradually across components, with no clear elbow or dominant latent dimension. The relatively low variance captured by the leading components indicates that semantic information is widely distributed across many dimensions rather than concentrated in a small number of dominant factors. This diffuse variance structure suggests that visitor experiences are encoded as complex and multidimensional constructs that cannot be adequately summarized by a few latent axes. The two-dimensional PCA projection (Figure 4) further illustrates this continuity, showing substantial overlap between polarity-labeled reviews rather than clear separation. A similar pattern is observed in the UMAP projection (Figure 5), where local neighborhood structure is preserved, but global categorical boundaries remain indistinct. Consequently, low-dimensional projections should be interpreted as exploratory visual aids rather than definitive representations of experiential structure.

Explained variance ratios obtained from PCA applied to the normalized embedding matrix.

Table 5. Principal Component Analysis of the Semantic Embedding Space.

Component	Explained Variance (%)	Cumulative Variance (%)
PC1	7.19	7.19
PC2	5.61	12.80
PC3	4.34	17.14
PC4	3.46	20.61
PC5	3.12	23.73

**Figure 2.** PCA scree plot of the semantic embedding space (first 50 components).**Figure 3.** Cumulative variance explained by the first 50 principal components.

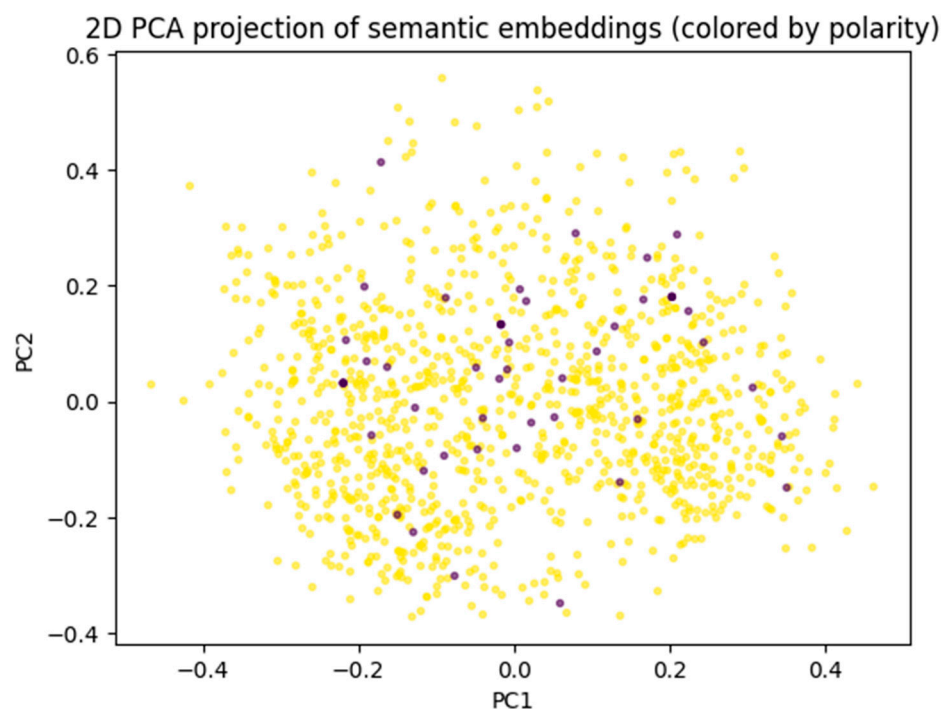


Figure 4. Two-dimensional PCA projection of the semantic embedding space with polarity labels.

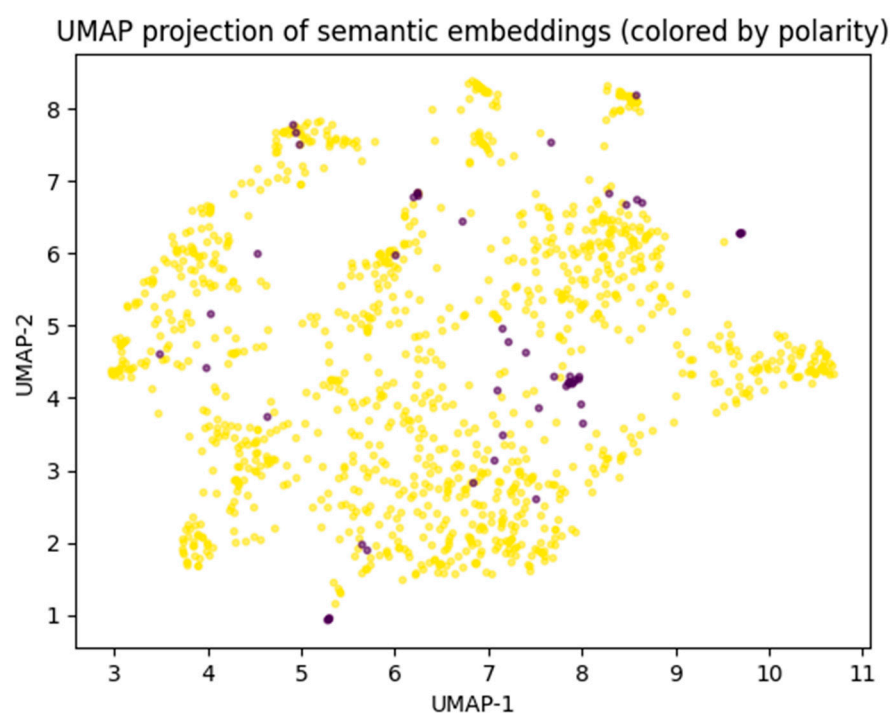


Figure 5. Two-dimensional UMAP projection of the semantic embedding space with polarity labels.

4.5. Clustering Structure and Continuity of Visitor Experience

Table 6 presents the clustering validity metrics for the semantic embedding space. Silhouette values are consistently low across all configurations, ranging from 0.036 to 0.055. Although positive, these values indicate weak cluster separation, pointing to strong overlap among semantic groups. At the same time, SSE decreases as the number of clusters increases, as expected for the k-means algorithm. Taken together, these results suggest that the semantic space of reviews does not organize into clearly defined clusters. Instead, visitor experiences are distributed along continuous and overlapping semantic dimensions. This

means that there are no clean, distinct types of experiences (e.g., “satisfied,” “dissatisfied,” “adventure-oriented,” “family-oriented”), but rather gradations of meaning characterized by proximity, distance, and gradual transitions between reviews. These findings indicate that forcing discrete cluster solutions onto the embedding space risks oversimplifying the underlying semantic structure. Rather than reflecting distinct experiential categories, the observed patterns are more consistent with continuous semantic gradients shaped by overlapping motivations, perceptions, and contextual factors. This result directly supports the study’s central argument that semantic embeddings should be examined as continuous metric spaces prior to any segmentation or task-specific modeling.

Table 6. Clustering Validity Metrics for the Semantic Embedding Space.

Number of Clusters (k)	Silhouette Score	Within-Cluster SSE
2	0.055	601.33
3	0.048	583.57
4	0.041	569.82
5	0.036	559.13
6	0.037	550.02
7	0.036	542.01
8	0.037	535.89
9	0.039	528.46
10	0.041	521.64

Note: $k = 1$ is not reported because silhouette statistics are undefined for a single-cluster solution.

5. Discussion

5.1. Interpretation of Semantic Structure in Visitor Experience

The results of this study demonstrate that the internal organization of transformer-based semantic embeddings does not align with how their outputs are commonly interpreted in applied research and practice. In tourism studies and protected area management, semantic embeddings are frequently used to segment visitor experiences into stable categories or types. The findings presented here challenge this assumption by showing that visitor experiences in national parks are not naturally organized into clearly separable groups within the embedding space. Empirical analysis reveals that positive and negative evaluations, appreciation of natural landscapes, and criticism of infrastructure or management frequently coexist within the same textual narratives. Rather than simplifying or resolving this complexity, transformer-based models preserve it as continuous, overlapping semantic configurations. The absence of strong cluster separation and the weak explanatory power of low-dimensional projections indicate that experiential meaning is distributed across gradual semantic gradients rather than discrete experiential classes.

Misinterpretation arises when analytical techniques that assume rigid boundaries—such as hard clustering or categorical segmentation—are imposed on inherently continuous semantic spaces. While such approaches may yield apparently clear results, they introduce artificial structure that does not reflect the underlying organization of visitor discourse. As a result, analytical clarity is achieved at the expense of experiential nuance. This mismatch between data structure and interpretative practice represents a central empirical insight of the study.

5.2. Theoretical Implications

From a theoretical perspective, the findings contribute to ongoing debates in tourism and experience research by questioning frameworks that conceptualize visitor experience as stable, discrete, and classifiable. The observed semantic continuity supports theoretical views that treat experience as relational, fluid, and context-dependent rather than as a fixed set of attributes or types. The study shows that experiential meaning emerges through overlapping evaluations and situational trade-offs rather than through clearly defined experiential states. By demonstrating that semantic embeddings encode ambivalence, simultaneity, and transition rather than categorical separation, the results call into question the theoretical validity of rigid segmentation models when applied to experiential narratives. This contributes to a more nuanced understanding of the visitor experience as dynamically constructed through the interaction among environmental conditions, individual expectations, and situational context.

5.3. Methodological Implications

Methodologically, the study highlights the need to reconsider how transformer-based semantic embeddings are used in quantitative text analysis. Embeddings are often treated as neutral feature generators, with their validity inferred from downstream predictive performance. The results presented here demonstrate that such an assumption is insufficient, as the internal numerical and geometric properties of embedding spaces strongly shape analytical outcomes. The analysis shows that similarity distributions, variance concentration, and clustering behavior provide critical diagnostic information about how meaning is structured within the embedding space. When these properties are ignored, common analytical practices—such as unsupervised clustering or low-dimensional visualization—risk imposing artificial order on continuous semantic structures. This can lead to overinterpretation of weak patterns and to unwarranted claims about the existence of distinct experiential groups.

The proposed framework, therefore, contributes a methodological layer that precedes task-specific modeling. By examining numerical stability, semantic continuity, and spatial organization before applying predictive or classificatory models, researchers can better assess whether their analytical tools are appropriate to the data's structure. This approach promotes more transparent, cautious, and structurally informed use of AI-based semantic models.

5.4. Practical Implications for Protected Area Management

The findings have direct implications for the use of textual analytics in national park management and similar applied contexts. Visitor experience management is often framed in terms of identifying satisfied and dissatisfied groups or segmenting visitors into distinct categories. The results suggest that such an approach may overlook important transitional and mixed experiences that are distributed across the semantic space. These transitional experiences often function as early indicators of emerging issues, such as crowding pressures, accessibility conflicts, or mismatches between visitor expectations and actual conditions. Analytical frameworks that rely exclusively on categorical interpretation may fail to detect such signals. By contrast, examining semantic continuity and overlap allows for a more sensitive understanding of experiential dynamics and supports more adaptive and responsive management strategies.

5.5. Limitations and Future Research Directions

Several limitations of the study should be acknowledged. First, the empirical analysis is based on a single corpus of English-language national park reviews, which limits direct

generalization to other languages, cultural contexts, or forms of experiential discourse. Second, the study focuses on a specific class of transformer-based sentence embeddings; alternative architectures or dimensionalities may exhibit different geometric properties. Third, the analysis is exploratory and diagnostic in nature and does not assess causal relationships or predictive accuracy. Future research could extend the proposed framework across different domains, languages, and embedding models to evaluate the robustness of the observed structural properties. Longitudinal datasets could be used to examine how semantic spaces evolve over time, while comparative studies could explore how different embedding architectures shape experiential representation.

6. Conclusions

This study contributes to the quantitative analysis of unstructured textual data by advancing a methodological perspective that treats transformer-based semantic embeddings as analyzable high-dimensional metric spaces rather than as opaque inputs to downstream predictive models. By systematically examining numerical stability, similarity structure, variance distribution, and clustering behavior, the research demonstrates that semantic embeddings encode experiential meaning through continuous and overlapping configurations rather than through, overlapping configurations rather than discrete, naturally separable categories. The findings highlight that strong predictive performance alone is not sufficient to justify the analytical use of semantic embeddings in contexts where interpretability and decision relevance are critical. Instead, explicit examination of the internal numerical and geometric properties of embedding spaces is necessary to ensure that subsequent analyses are grounded in the actual structure of the data. The proposed framework provides a generalizable approach for conducting such diagnostics prior to task-specific modeling.

From an applied perspective, the study underscores the importance of cautious interpretation when using AI-based text analytics to support decision-making in national park management and related domains. Simplified categorical representations of visitor experience may obscure transitional, mixed, or emerging patterns that are essential for adaptive and sustainable management. Integrating representational diagnostics with task-oriented analytics enables more nuanced and responsible use of visitor-generated textual data. The framework introduced in this study is not limited to tourism research and can be applied across domains where large-scale unstructured text is used to inform quantitative analysis. Future research may extend this approach to alternative embedding architectures, multilingual corpora, and longitudinal datasets in order to further assess the robustness and generality of the observed structural properties. By shifting attention from model outputs to the structure of meaning within representations, this study supports more interpretable, reliable, and accountable integration of AI-based text analytics into empirical research and practice.

Author Contributions: Conceptualization, A.V. and D.K.; methodology, A.V. and G.P.; software, D.K. and D.S.; validation, A.V., D.K. and V.M.; formal analysis, D.K. and A.V.; investigation, D.K., G.P. and V.M.; resources, A.V. and D.K.; data curation, A.V. and G.P.; writing—original draft preparation, A.V. and D.S.; writing—review and editing, D.K., G.P. and V.M.; visualization, A.V. and D.S.; supervision, D.K., A.V. and V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset is publicly available on Kaggle (Parks Reviews Dataset). Derived embeddings and analysis scripts are available from the corresponding author upon reasonable request.

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

References

1. Zachlod, C.; Samuel, O.; Ochsner, A.; Werthmüller, S. Analytics of social media data—State of characteristics and application. *J. Bus. Res.* **2022**, *144*, 1064–1076. [\[CrossRef\]](#)
2. Li, S.; Liu, F.; Zhang, Y.; Zhu, B.; Zhu, H.; Yu, Z. Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. *Mathematics* **2022**, *10*, 3554. [\[CrossRef\]](#)
3. Grljević, O. Topic modeling in hospitality and tourism research: Application areas, business insights, and managerial implications. *Hotel Tour. Manag.* **2025**, *13*, 137–153. [\[CrossRef\]](#)
4. Mirčetić, V.; Popović, G.; Vukotić, S. Unveiling the characteristics of the EU charismatic leaders using PIPRECIA-S method. *J. Process Manag. New Technol.* **2024**, *12*, 99–109. [\[CrossRef\]](#)
5. Lee, J.; Song, C.H. From Unstructured Feedback to Structured Insight: An LLM-Driven Approach to Value Proposition Modeling. *Electronics* **2025**, *14*, 4407. [\[CrossRef\]](#)
6. Greco, G.; Boch, T.; Fernique, P.; Marchand, M.; Allen, M.; Pineau, F.-X.; Baumann, M.; Molinaro, M.; De Pietri, R.; Branchesi, M.; et al. Encapsulating textual contents into a MOC data structure for advanced applications. *Astron. Comput.* **2026**, *54*, 101014. [\[CrossRef\]](#)
7. Teles, A.S.; de Moura, I.R.; Silva, F.; Roberts, A.; Stahl, D. EHR-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. *Inf. Fusion* **2025**, *118*, 102981. [\[CrossRef\]](#)
8. Blundo, C.; Cimato, S. A Bag of Words Model for Efficient Discovery of Roles in Access Control Systems. *Comput. Secur.* **2025**, *162*, 104808. [\[CrossRef\]](#)
9. Schreiber, M.; Jenny, G.J.; Hürlimann, M.; Parfenova, Y.; von Däniken, P.; Cieliebak, M. A discourse on the use of machine learning (ML) in personality psychology: Can we expect ML to predict questionnaire scores from idiographic text-based data? *J. Res. Personal.* **2025**, *119*, 104666. [\[CrossRef\]](#)
10. Zhou, J.; Ye, Z.; Zhang, S.; Geng, Z.; Han, N.; Yang, T. Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data. *Heliyon* **2024**, *10*, e35945. [\[CrossRef\]](#)
11. Patil, R.; Boit, S.; Gudivada, V.; Nandigam, J. A survey of text representation and embedding techniques in nlp. *IEEE Access* **2023**, *11*, 36120–36146. [\[CrossRef\]](#)
12. Redwan, K.; Datto, S.; Ahmed, M.; Masum, H.R.; Al Sohan, M.F.A.; Shufian, A. A multimodal deep learning framework for integrating visual, textual and categorical features in retail price estimation. *Array* **2025**, *28*, 100565. [\[CrossRef\]](#)
13. Yang, C.; Zhang, Y. Public emotions and visual perception of the East Coast Park in Singapore: A deep learning method using social media data. *Urban For. Urban Green.* **2024**, *94*, 128285. [\[CrossRef\]](#)
14. Mirčetić, V.; Mihić, M. Smart Tourism as a Strategic Response to Challenges of Tourism in the Post-COVID. Sustainable Business Management and Digital Transformation: Challenges and Opportunities in the Post-COVID Era. *Lect. Notes Netw. Syst. Springer* **2022**, *562*, 445–463. [\[CrossRef\]](#)
15. Yao, Z.; Bao, Y.; Liu, X.; Shan, Y.; He, T. Vibration and noise reduction characteristics of double-layer stiffened plates embedded with acoustic black holes. *Appl. Acoust.* **2025**, *237*, 110767. [\[CrossRef\]](#)
16. Pietrasik, M.; Reformat, M.Z. Probabilistic Coarsening for Knowledge Graph Embeddings. *Axioms* **2023**, *12*, 275. [\[CrossRef\]](#)
17. Mellina-Andreu, J.L.; Cisterna-García, A.; Botía, J.A. Data-driven interpretation of dimensions in an embedding language model based on a reference knowledge graph. *Knowl. Based Syst.* **2025**, *330*, 114507. [\[CrossRef\]](#)
18. Du, K.-L.; Zhang, R.; Jiang, B.; Zeng, J.; Lu, J. Understanding Machine Learning Principles: Learning, Inference, Generalization, and Computational Learning Theory. *Mathematics* **2025**, *13*, 451. [\[CrossRef\]](#)
19. Jiang, S.; Chin, K.-S.; Qu, G.; Tsui, K.L. An integrated machine learning framework for hospital readmission prediction. *Knowl. Based Syst.* **2018**, *146*, 73–90. [\[CrossRef\]](#)
20. Ferhath, A.A. Machine learning: Techniques, applications, and metrics for enhanced vehicle performance. *J. Process Manag. New Technol.* **2025**, *13*, 1–11. [\[CrossRef\]](#)
21. Dinh, T.; Wong, H.; Lisik, D.; Koren, M.; Tran, D.; Yu, P.S.; Torres-Sospedra, J. Data clustering: A fundamental method in data science and management. *Data Sci. Manag.* **2025**; *in press*. [\[CrossRef\]](#)
22. Chen, A.; Chen, H.; Zhang, Z.; Yang, M.; Chen, Y.-Y. EmbTCN-Transformer: An Embedding Temporal Convolutional Network–Transformer Model for Multi-Trajectory Prediction. *Mathematics* **2025**, *13*, 3306. [\[CrossRef\]](#)
23. Hong, S.-K.; Jang, J.-S.; Kwon, H.-Y. Enhancing performance of transformer-based models in natural language understanding through word importance embedding. *Knowl. Based Syst.* **2024**, *304*, 112404. [\[CrossRef\]](#)
24. Noori, A.; Balafar, M.A.; Bouyer, A.; Salmani, K. Contrastive learning with transformers for meta-path-free heterogeneous graph embedding. *Appl. Soft Comput.* **2026**, *188*, 114506. [\[CrossRef\]](#)
25. Ho, H.-T.; Nguyen, T.-T.-D.; Le, N.Q.K.; Ou, Y.-Y. FAD-BERT: Improved prediction of FAD binding sites using pre-training of deep bidirectional transformers. *Comput. Biol. Med.* **2021**, *131*, 104258. [\[CrossRef\]](#)
26. Castro, A.P.; Wainer, G.A.; Calixto, W.P. Weighting construction by bag-of-words with similarity-learning and supervised training for classification models in court text documents. *Appl. Soft Comput.* **2022**, *124*, 108987. [\[CrossRef\]](#)

27. Sarıtaş, K.; Öz, C.A.; Güngör, T. A comprehensive analysis of static word embeddings for Turkish. *Expert Syst. Appl.* **2024**, *252*, 124123. [\[CrossRef\]](#)
28. Khan, H.M.; Basheer, S.; Quasim, M.T.; Al-Naimi, R.; Varadarajan, V.; Khan, A. A transformer-based deep learning framework with semantic encoding and syntax-aware LSTM for fake electronic news detection. *Comput. Mater. Contin.* **2025**, *86*, 1–25. [\[CrossRef\]](#)
29. Huang, S.; Chen, J.; Yu, C.; Li, D.; Zhou, Q.; Liu, S. DE-ESD: Dual encoder-based entity synonym discovery using pre-trained contextual embeddings. *Expert Syst. Appl.* **2025**, *276*, 127102. [\[CrossRef\]](#)
30. Zhao, W.; Wang, W. Investigating the performance of DistilBERT and LSTM-CNN models with GloVe embeddings for emotion detection from textual data. *Egypt. Inform. J.* **2025**, *32*, 100808. [\[CrossRef\]](#)
31. Mao, C.; Shuang, K.; Guo, J.; Qian, B.; Yang, Y.; Li, H. Cognition-aligned frequency filtering for sentence embeddings. *Inf. Process. Manag.* **2026**, *63*, 104415. [\[CrossRef\]](#)
32. Ohams, C.; Nair, S.; Bhattasali, S.; Resnik, P. A predictive coding model for online sentence processing. *J. Mem. Lang.* **2026**, *146*, 104705. [\[CrossRef\]](#)
33. Matsumoto, N.; Iijima, Y.; Lin, M.; Nishiguchi, Y.; Takano, K.; Raes, F. Semantic similarity among autobiographical memories is associated with rumination. *J. Behav. Ther. Exp. Psychiatry* **2026**, *90*, 102072. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Saoualih, A.; Perkumienė, D.; Safaa, L.; Škema, M.; Aleinikovas, M. Computational mining of empirical literature on forest recreation: A semantic-driven topic modeling approach based on advanced contextual embeddings. *Trees For. People* **2025**, *20*, 100877. [\[CrossRef\]](#)
35. Fu, X.; Liu, X.; Li, Z. Catching eyes of social media wanderers: How pictorial and textual cues in visitor-generated content shape users' cognitive-affective psychology. *Tour. Manag.* **2024**, *100*, 104815. [\[CrossRef\]](#)
36. Gao, J.; Yu, H.; Cheung, Y.-M.; Cao, J.; Wong, R.C.-W.; Zhang, Y. Shaping pre-trained language models for task-specific embedding generation via consistency calibration. *Neural Netw.* **2025**, *191*, 107754. [\[CrossRef\]](#)
37. Stanujkić, D.; Karabašević, D.; Popović, G.; Zavadskas, E.K.; Saračević, M.; Stanimirović, P.S.; Ulutaş, A.; Katsikis, V.N.; Meidute-Kavaliauskienė, I. Comparative Analysis of the Simple WISP and Some Prominent MCDM Methods: A Python Approach. *Axioms* **2021**, *10*, 347. [\[CrossRef\]](#)
38. Ettinger, A. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 34–48. [\[CrossRef\]](#)
39. Aggarwal, C.C.; Hinneburg, A.; Keim, D.A. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001*; Van den Bussche, J., Vianu, V., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2001; Volume 1973, pp. 420–434. [\[CrossRef\]](#)
40. Caronia, L.; Ranzani, F.; Benericetti, G.; Scattolini, C.; Chierigato, A. The visitors' book as a family-centered care tool: A corpus-based, multi-site study on the implementation of a narrative care practice in ICU. *Intensive Crit. Care Nurs.* **2026**, *92*, 104188. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Qiu, L.; Shen, L.; Liu, L.; Liu, J.; Chen, Y.; Xing, L. ST-NeRP: Spatial-temporal neural representation learning with prior embedding for patient-specific imaging study. *Comput. Biol. Med.* **2025**, *198*, 111266. [\[CrossRef\]](#)
42. Nivarthi, C.P.; Huang, Z.; Gruhl, C.; Sick, B. TRACE: Time series representation learning with contrastive embeddings for anomaly detection in photovoltaic systems. *Energy AI* **2025**, *23*, 100670. [\[CrossRef\]](#)
43. Pandhi, S.; Kumar, A. Assessing the influence of extraction techniques on the phytochemical composition of green coffee (*Coffea arabica*) using principal component analysis (PCA) and hierarchical cluster analysis (HCA). *J. Indian Chem. Soc.* **2025**, *102*, 102111. [\[CrossRef\]](#)
44. Teng, J. Financial data reduction and information retention strategy based on principal component analysis (PCA) algorithm. *Procedia Comput. Sci.* **2025**, *262*, 218–226. [\[CrossRef\]](#)
45. Orea-Giner, A.; Fuentes-Moraleda, L.; Villacé-Molinero, T.; Muñoz-Mazón, A.; Calero-Sanz, J. Does the implementation of robots in hotels influence the overall TripAdvisor rating? A text mining analysis from the Industry 5.0 approach. *Tour. Manag.* **2022**, *93*, 104586. [\[CrossRef\]](#)
46. Tunca, S.; Balcioglu, Y.S. Mapping digital satisfaction dimensions in mobile fashion retail: Service-Dominant Logic in the Turkish market. *J. Retail. Consum. Serv.* **2026**, *88*, 104530. [\[CrossRef\]](#)
47. Wani, A.A. Comprehensive review of dimensionality reduction algorithms: Challenges, limitations, and innovative solutions. *PeerJ Comput. Sci.* **2025**, *11*, e3025. [\[CrossRef\]](#)
48. Colace, F.; Gaeta, R.; Lorusso, A.; Pellegrino, M.; Santaniello, D. New AI challenges for cultural heritage protection: A general overview. *J. Cult. Herit.* **2025**, *75*, 168–193. [\[CrossRef\]](#)
49. Tkaczynski, A.; Rundle-Thiele, S.R.; Beaumont, N. Segmentation: A tourism stakeholder view. *Tour. Manag.* **2009**, *30*, 169–175. [\[CrossRef\]](#)
50. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, *46*, 243–256. [\[CrossRef\]](#)

51. Wiroomsri, N.; Preedasawakul, O. A correlation-based fuzzy cluster validity index with secondary options detector. *Fuzzy Sets Syst.* **2026**, *523*, 109632. [[CrossRef](#)]
52. Kamat, M.; Jagasia, J.; Vaidya, A.; Surve, O. Embedding-based decision support framework for large-scale content analysis. *Knowl. Based Syst.* **2026**, *332*, 114926. [[CrossRef](#)]
53. Chang, Y.-T.; Chen, S.-F. Sensory-CoKGE: A contextualized knowledge graph embedding framework using language models for converting text-based food attributes into numerical representation. *Expert Syst. Appl.* **2026**, *299*, 130191. [[CrossRef](#)]
54. Greco, C.; Ianni, M. A formal framework for LLM-assisted automated generation of Zeek signatures from binary artifacts. *Future Gener. Comput. Syst.* **2026**, *175*, 108086. [[CrossRef](#)]
55. Wu, H.; He, B.; Xie, D.; Chen, C.; Zhang, W. Self-spectacularization of tourists in visual social media: A computer vision and deep learning approach to socio-cultural body schemas. *Chaos Solitons Fractals* **2025**, *200*, 117098. [[CrossRef](#)]
56. Shomoye, M.; Zhao, R. Automated emotion recognition of students in virtual reality classrooms. *Comput. Educ. X Real.* **2024**, *5*, 100082. [[CrossRef](#)]
57. Zhao, J.; Geipel, J. Super-resolve satellite imagery to perform on par with UAV-borne hyperspectral imagery in predicting spring wheat physiological parameters using transformer models. *Comput. Electron. Agric.* **2026**, *240*, 111204. [[CrossRef](#)]
58. Larsson, C. *5G Networks: Planning, Design, and Optimization*; Elsevier: Amsterdam, The Netherlands, 2018. [[CrossRef](#)]
59. Orkphol, K.; Yang, W. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet* **2019**, *11*, 114. [[CrossRef](#)]
60. McKenzie, J.S.; Donarski, J.A.; Wilson, J.C.; Charlton, A.J. Analysis of complex mixtures using high-resolution nuclear magnetic resonance spectroscopy and chemometrics. *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *59*, 336–359. [[CrossRef](#)]
61. Palacio-Niño, J.O.; Berzal, F. Evaluation metrics for unsupervised learning algorithms. *arXiv* **2019**, arXiv:1905.05667. [[CrossRef](#)]
62. Yahyaoui, H.; Own, H.S. Unsupervised clustering of service performance behaviors. *Inf. Sci.* **2018**, *422*, 558–571. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.