technologies

MDPI

Article

Semantic Diversity and Visitor Sentiments in Ecotourism Using Geospatial Natural Language Processing of Social Sensing Data

Asamaporn Sitthi [1,*], Uday Pimple [2], Pattamaporn Wongwiriya [3] and Can Trong Nguyen [4]

[1] Department of Geography, Faculty of Social Sciences, Srinakharinwirot University, 114 Sukhumvit 23, Wattana District, Bangkok 10110, Thailand
[2] The Joint Graduate School of Energy and Environment and Centre of Excellence on Energy Technology and Environment, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand; upimple@gmail.com or uday.pim@kmutt.ac.th
[3] Urban and Regional Planning Program, Faculty of Architecture, Khon Kaen University, Khon Kaen 40002, Thailand; pattamawong@kku.ac.th
[4] Environment Centre, Charles University, 16200 Prague, Czech Republic; can.nguyen@czp.cuni.cz
* Correspondence: asamaporn@g.swu.ac.th or as.sitthi@gmail.com

Abstract

This study aimed to investigate visitors' perceptions of ecotourism landscapes in Thailand's protected areas by integrating spatial, semantic, and affective information derived from social-sensing textual data. To this end, it examined the influence of ecosystem characteristics on thematic expressions and emotional tones across five national parks. Methodologically, a spatio-semantic, natural language processing (NLP) social-sensing framework was developed using geotagged Flickr tags and YouTube comments from international (English) and domestic (Thai) tourists. Textual data were preprocessed (cleaned, normalized, and tokenized) and analyzed using latent Dirichlet allocation (LDA), sentiment analysis, and diversity metrics. Concurrently, YouTube comments were similarly processed using LDA and rule-based sentiment analysis. Subsequently, a Diversity × Topic × Season matrix integrated Flickr-derived indicators with YouTube-derived sentiment and topic dominance. Binary logistic regression was applied to examine cross-platform relationships. The analysis identified three dominant themes, namely Nature & Landscapes, Travel & Activities, and Feelings & Experiences. Forest-mountain parks showed high semantic diversity and strong positive sentiment, whereas marine parks exhibited narrower but predominantly positive activity-driven discourses. Moreover, seasonal variation was evident, with summer and winter yielding the highest diversity and positivity. Building on these results, this study devised a spatio-semantic framework for analyzing variation in visitor expressions across parks and seasons. It captures eco-awareness, perceptions, and preferred activities in protected landscapes. From a practical perspective, semantic diversity and sentiment indicators can support ecotourism management through improved visitor monitoring and communication strategies.

Keywords: social sensing; ecotourism perception; sentiment analysis; natural language processing; topic modeling; spatio-semantic framework

Academic Editors: Tamara Gajić and Pedro Antonio Gutiérrez

Received: 14 May 2026
Revised: 15 July 2026
Accepted: 21 July 2026
Published: 23 July 2026

1. Introduction

Tourism encompasses a wide range of visitor preferences, ranging from well-known attractions to lesser-known cultural and natural settings, offering seasonal experiences such as trekking, camping, birdwatching, and diving across diverse ecosystems [1,2]. Although tourism is a major global industry, its environmental impacts vary across countries, depending on regional and country-specific policies, and these impacts remain insufficiently understood. Southeast Asian countries offer year-round tourism owing to their tropical climates and favorable temperatures. Thailand is one of the most popular destinations and has developed into a major tourism hub, driving economic growth while providing additional income to local people and communities [3].

Accordingly, the Five-Year Development Plan (2023–2027) of the Department of National Parks, Wildlife and Plant Conservation (DNP) outlines an implementation strategy focused on developing a resilient tourism ecosystem, promoting eco-friendly tourism, and ensuring the sustainable use of ecosystem services [3]. The plan emphasizes environmentally responsible tourism and the adoption of digital innovation to improve monitoring, visitor engagement, and adaptive management in protected areas.

Over the past decade, increasing awareness of the importance of forests, oceans, and environmental sustainability has reshaped tourism patterns. This shift has led to a growing interest in nature-based and sustainable travel forms, particularly ecotourism. Ecotourists are particularly attracted to national parks because they value natural landscapes and biodiversity while seeking environmentally responsible recreational experiences [4]. Ecotourism therefore serves a dual role: supporting biodiversity conservation and promoting environmental awareness among visitors and local communities [5,6].

To effectively implement these strategies, an understanding of visitors' emotional and experiential connections to natural environments, including tourist perceptions and cognitive engagement, is needed. Tourists' perceptions of natural landscapes are shaped by the interactions between natural features and human factors, such as cognitive and emotional responses. This relationship has been widely examined in environmental perception and landscape aesthetics research. According to information processing theory [7], complexity and mystery trigger cognitive engagement, which is discrepant and subjective in nature. In a tourism context, individual perceptions are often expressed in the language used by visitors to describe their experiences [8].

The concept of semantic diversity provides a perspective for understanding how visitors linguistically express their perceptions of tourism destinations. Visitor perceptions of natural landscapes are shaped by cognitive and emotional responses to environmental features [9]. This diversity reflects the range of environmental perceptions, indicating cognitive richness in how visitors perceive environmental attributes and experiential complexity [10–12]. Ecologically diverse environments, such as protected areas with varied habitats, species, and landscape features, tend to stimulate a wider range of visitor perceptions and experiences [13]. These may be reflected in more diverse linguistic expressions in visitor narratives. Such expressions offer insight into how individuals perceive, interpret, and emotionally respond to landscapes, aligning with environmental psychology and tourism experience frameworks that highlight the role of cognitive and affective responses in shaping destination perceptions.

Traditionally, authorities have relied on field-based methods such as interviews, focus group discussions, and online questionnaires. However, these approaches are often time-consuming and cannot provide continuously updated data for monitoring and assessing tourist behavior [14]. Advancements in communication technologies have transformed information into a crucial digital resource through social media platforms [5,15–19]. Among these, Flickr and YouTube have particularly exceptional complementary strengths. Flickr captures the essence of a place through images of landscape beauty and

https://doi.org/10.3390/technologies14080456

distinctiveness, accompanied by detailed narratives [14]. Meanwhile, YouTube offers audiovisual storytelling and audience interaction through comments, providing deeper insight into how visitors perceive destinations. Large-scale user-generated content significantly influences visitor decision-making processes. It shapes destination preferences and encourages tourists to immerse themselves in the atmospheres portrayed by others. It also provides valuable information for stakeholders, such as tourism operators, government agencies, the corporate sector, and local communities [20], enabling them to tailor their strategies to maximize benefits while minimizing environmental impacts.

Leveraging this information requires NLP to transform text into actionable knowledge. These techniques enable the analysis of word frequencies, clustering patterns, and sentiment. Previous studies have employed approaches ranging from lexicon-based methods in deep learning models to interpret visitor emotions and experiences [17,19,21–23]. Lexicon-based techniques rely on sentiment dictionaries, such as SentiWordNet, VADER, or custom-built lists, to assign polarity scores to text [24]. Consequently, an NLP-driven sentiment analysis applied to social sensing datasets offers a robust means of assessing tourist behaviors and perceptions in ecotourism contexts [17,20,25,26].

A key challenge in analyzing such social media-based data lies in the inherent complexity of language, which varies across contexts, including linguistic variations and stylistic features. While previous studies have combined social sensing, topic modeling, and sentiment analysis in urban tourism or general destination contexts [22,27–30], the systematic application within Thailand's protected areas, integrating park-level ecological heterogeneity, seasonal variation, and multi-platform social sensing remains limited. This study addresses this gap by integrating geotagged Flickr images and YouTube comments to examine how visitors perceive and emotionally engage with Thailand's national parks. Using NLP-based topic modeling and sentiment analysis, spatial and semantic indicators are linked to reveal seasonal patterns and thematic diversity. The contribution of this study lies in three areas of integration: (1) the simultaneous use of Flickr and YouTube as complementary spatial and affective data sources within protected-area contexts; (2) the operationalization of Shannon entropy as a semantic diversity indicator bridging ecological and linguistic complexity; and (3) the application of a Diversity × Topic × Season framework to reveal how landscape type and seasonality jointly shape visitor sentiment.

2. Study Area and Dataset

2.1. Study Area

This study focused on five of Thailand's most popular national parks, selected to represent a range of ecosystems, including coastal, forest, and mountain environments. This ecological variety captures the necessary landscape heterogeneity to investigate how environmental variation influences semantic diversity and visitor sentiment. This includes marine ecosystems: Koh Chang and Similan National Parks (coral reefs, diving); Forest Ecosystems: Khao Yai National Park (biodiversity, nature trails); and Mountain Ecosystems: Doi Inthanon and Phu Kradueng National Parks (hiking and adventure). The inclusion of these five national parks underscores Thailand's environmental diversity and provides a foundation for examining relationships between landscape characteristics and tourist perceptions, as well as addressing the challenges of promoting sustainable ecotourism management across protected areas (Figure 1; [3]).

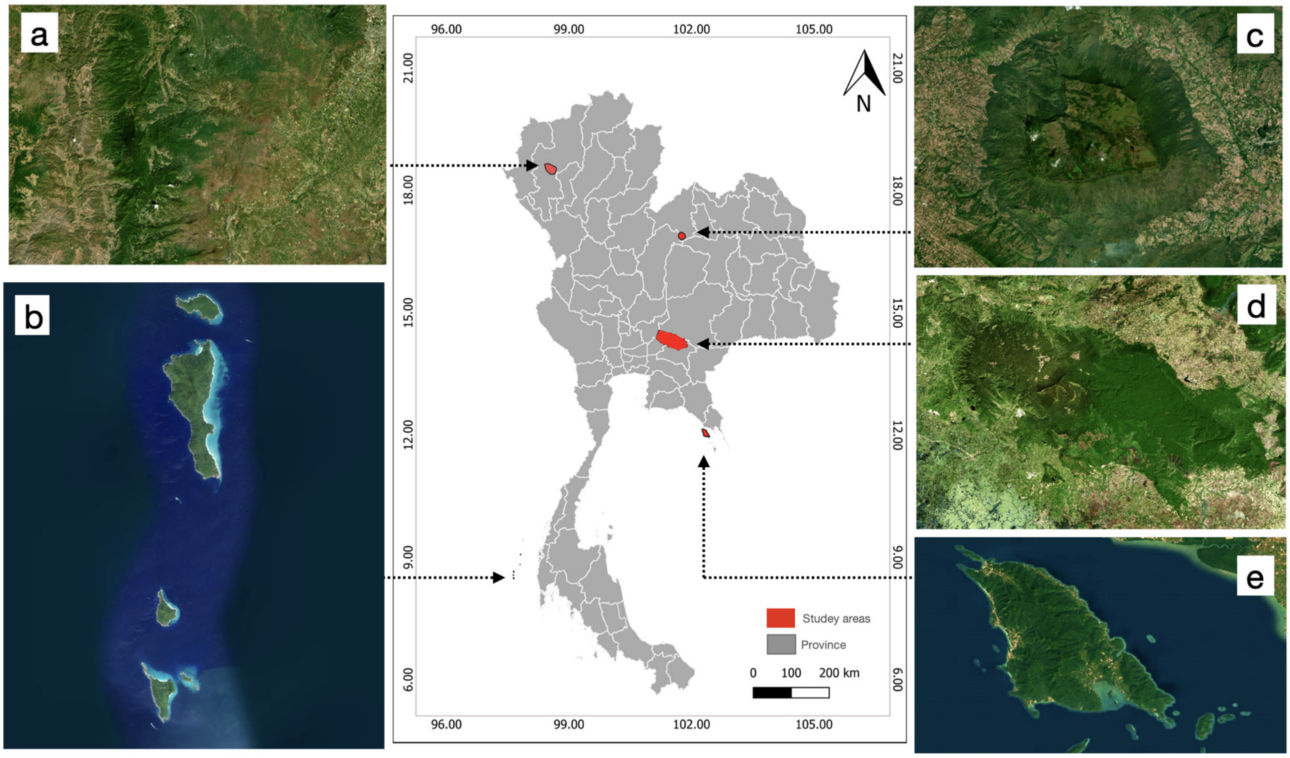

Figure 1. Locations of the five selected national parks. (**a**) Doi Inthanon National Park, (**b**) Similan Islands, (**c**) Phu Kradueng National Park, (**d**) Khao Yai National Park, and (**e**) Koh Chang. Note: Areas marked in red on the map indicate the location of the national parks.

2.2. Dataset

Flickr and YouTube were used as data sources owing to the following three criteria: spatial metadata availability, linguistic complementarity, and open Application Programming Interface (API) access. Flickr's geotagged API provides coordinates for each photograph, enabling fine-grained spatial analysis that is not feasible with Instagram (which deprecated its location-based API in 2020) or TikTok. YouTube's public comment API allows extraction of time-stamped, park-specific visitor narratives. Although Twitter/X offers real-time data, posts therein are rarely geotagged to national park boundaries and tend to be reactive rather than experiential. The platform choice inevitably conditions the results: Flickr predominantly reflects international, photography-oriented visitors, whereas YouTube primarily reflects domestic audiences. These biases are addressed in the platform-specific analysis and acknowledged as a study limitation.

Flickr's geotagged photographs offer detailed spatial information, reflecting where visitors are physically present and which landscapes attract their attention. In this study, geotagged data were collected from the Flickr platform using the photosearcher package [31] in R (version 2025.05.0+496/), which connects to the Flickr Application Programming Interface [32] via an authenticated key and secret. The search query was restricted to images associated with the keyword "Thailand" and the respective park name and constrained within the national bounding-box coordinates. To ensure spatial and temporal consistency, only photographs taken between 2018 and 2024 were included, covering pre-COVID, COVID-19, and post-pandemic phases. The geotagged parameter was applied to retrieve only images with valid geographic coordinates. Metadata for each photograph, including the title, description, tags, latitude, longitude, and date and time of capture, were downloaded. To obtain an overview of the spatial distribution of visitor activity, a density grid map was created to illustrate where photographs were most frequently taken. As expected, the highest concentrations occurred around Bangkok, Chiang Mai, Khao Yai, and popular eastern and southern coastal areas, while other regions showed comparatively fewer data points. These density patterns closely align with national park tourism statistics [4].

While Flickr photographs represent virtual presence, YouTube videos offer complementary insights into visitors' engagement and public interest. Textual data were retrieved from YouTube using the YouTube Data API v3 [33] through a custom Google Apps Script implementation. The API enabled the extraction of metadata, including the video title, description, tags, channel ID, and upload date. Queries were filtered for the five national parks using a combination of location-based filters and park-specific keywords. In addition to metadata, the API also provided engagement metrics, such as views, likes, comments, and channel subscriber counts, which capture the visibility and popularity of the locations. These indicators were incorporated to represent audience interaction and the deeper emotional layers of the visitor experience.

These data derived from online platforms may contain inherent biases related to user demographics, language use, and platform-specific behavior. For example, Flickr tends to attract photography-oriented international users, whereas YouTube often reflects more informational or domestic audiences. Furthermore, the spatial and temporal distribution of user-generated content is not uniform, which may lead to uneven data density across parks and seasons. These factors may influence the representativeness of observed patterns and should be considered when interpreting the results. Consequently, the collected data may primarily represent the perceptions of active social media users rather than the entire visitor population, limiting the generalizability of the findings to all visitors to protected areas. Nevertheless, combining multiple platforms and multilingual data helps mitigate some of these biases and provides a broader perspective on visitor perceptions in protected areas. Such integration also enables a more comprehensive examination of how online visitors engage with Thailand's national parks [34], while highlighting spatial and semantic differences in ecotourism across sites.

3. Methodology

The overall analytical workflow is summarized in Figure 2. Data were first collected from the Flickr and YouTube APIs using location- and keyword-based filters. After storing and cleaning the raw data, textual content was processed using NLP techniques, including tokenization, stopword removal, and lemmatization in both English and Thai. Bag-of-Words and Term Frequency–Inverse Document Frequency (TF–IDF) representations were then generated to extract textual features. Topic modeling, sentiment analysis, and Shannon entropy were applied to derive thematic categories, emotional polarity, and tag diversity, respectively. Cross-platform relationships were examined using binary logistic regression models. Finally, spatial visualization techniques were applied to map semantic and affective indicators, revealing the spatial patterns of visitor perceptions across the study sites.

https://doi.org/10.3390/technologies14080456

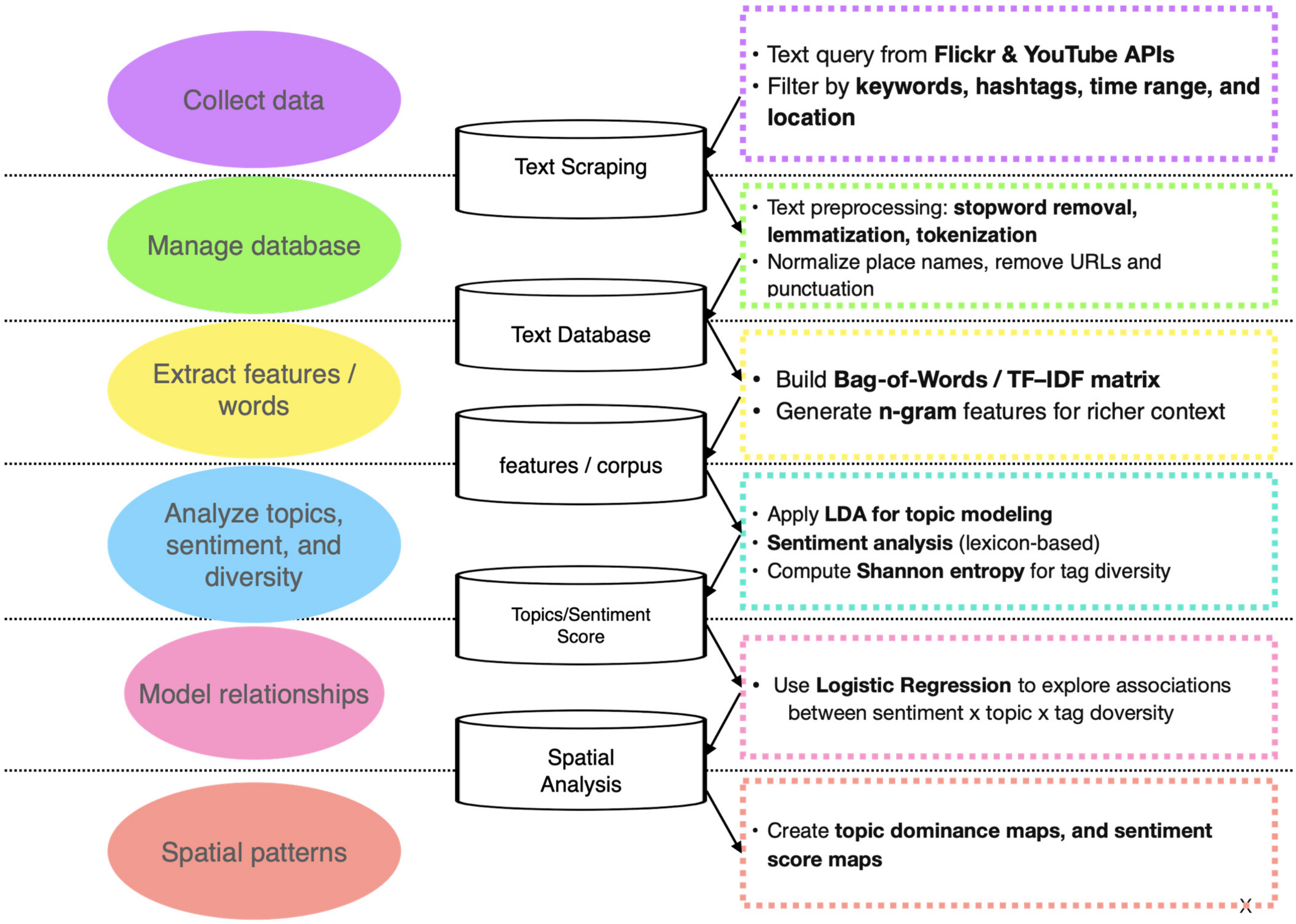

Figure 2. Workflow of the spatio–semantic analysis integrating Flickr tags and YouTube comments. Notes: TF–IDF = Term Frequency–Inverse Document Frequency; LDA = Latent Dirichlet Allocation; BoW = Bag-of-Words; n-gram = contiguous sequence of n words; API = Application Programming Interface.

3.1. Text Preprocessing

R packages, including tidyverse for data wrangling, stringr for text cleaning, quanteda for feature extraction (TF–IDF), topicmodels for LDA, and Vegan for computing Shannon entropy, were employed [35–38]. Python 3.11 was used with Natural Language Toolkit (NLTK) and PyThaiNLP for bilingual text preprocessing, scikit-learn for TF–IDF and feature extraction, gensim for topic modeling, and statsmodels for logistic regression analysis [39–43]. All text data from Flickr and YouTube, including titles, descriptions, tags, and comments, were processed using a standardized NLP preprocessing pipeline. The text was converted to lowercase, and irrelevant elements such as URLs, numeric strings, punctuation, emojis, and other non-alphabetic symbols were removed, retaining only English and Thai characters.

Consecutive white spaces were collapsed into a single space, and a custom dictionary was applied to normalize multi-word or variably spelled place names. For example, "Doi-inthanon" and "Doi-inthanon" were unified as "doiinthanon," while "Khao Yai" and "khao-yai" were standardized as "khaoyai".

The corpus was segmented into tokens. For English texts, tokenization was performed using regular expression matching in NLTK, extracting contiguous alphabetic sequences while discarding numbers and symbols. A custom stop-word list was developed to capture both general English function words (e.g., "the," "is," "are," "Photography," "Cannon," "Instagram," "uploaded," "picoftheday," and "follow"), and platform-spe-

https://doi.org/10.3390/technologies14080456

cific, high-frequency but semantically weak terms, commonly found on Flickr. However, for YouTube data, redundant informal, conversational, and low-information terms characteristic of user-generated Thai content were removed (e.g., "that," "krub," "ka," "very much," "ok," "like," "personally," "indeed," and "too"). Tokens shorter than two characters were also discarded. English comments were processed with NLTK, whereas Thai comments were processed using the newmm engine (PyThaiNLP) [40], which effectively handles continuous Thai script without explicit word delimiters.

Lemmatization was applied to English tokens using the NLTK WordNetLemmatizer to reduce inflectional forms to their base or dictionary forms. Thai tokens were not lemmatized due to the lack of reliable morphological simplification tools; instead, they were retained in their segmented form. This approach ensured that English words were normalized to canonical forms, while Thai words preserved their original, cleaned segmentation.

To capture frequent co-occurrences and descriptive expressions, unigram and bigram features were extracted using the scikit-learn CountVectorizer. Tokens were transformed into an n-gram representation, with English stopwords excluded during vectorization. The resulting frequency patterns highlighted iconic unigrams (e.g., "elephant," "species," "sunset," and "ocean") and bigrams (e.g., "national park," "scuba diving," "amphibian frog," "white-capped bird," and "bird watching"), providing insights into dominant eco-themes across Flickr and YouTube content [34].

3.2. Text Analysis

3.2.1. Topic Modeling Using Latent Dirichlet Allocation

Semantic topics were further explored across the national parks using Flickr's descriptive images such as scenery and landscapes, while the YouTube corpus reflected visitors' perceptions and experiences. The textual data were transformed into a document–term matrix using scikit-learn's CountVectorizer, which represents each document as a bag-of-words vector. Both unigrams and bigrams were extracted to capture not only individual keywords but also recurring multi-word expressions. LDA models were trained with different numbers of predefined topics (k = 3–10) [44]. Each model generated two probability distributions: the document–topic distribution, which describes the proportion of topics within each document, and the topic–word distribution, which defines the probability of words belonging to each topic. Three metrics were used for model evaluation: (1) perplexity, which assesses predictive performance; (2) topic coherence, which evaluates the semantic consistency of top words; and (3) topic diversity, which indicates the proportion of dominant words across topics, defined as

$$Perplexity\ (D_{test}) = exp\left(-\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d}\right) \quad (1)$$

$$Topic\ Diversity = \left|\frac{\cup_{k=1}^{K} W_k}{K^{*}N}\right| \quad (2)$$

Here, D_{test} is the held-out test corpus; w_d is the set of words in document d; N_d represents the number of words in document d; M is the total number of documents; and K is the number of topics. W_k refers to the top N words in topic k, and $\cup_{k=1}^{K} W_k$ indicates the count of distinct words across all topics. To ensure that the selected topic model effectively represented both Flickr landscape content and YouTube perspectives, candidate models were compared, and the optimal number of topics (k) was chosen by balancing high topic coherence, high topic diversity, and reasonable perplexity. Each document was assigned to a dominant topic based on the maximum posterior probability from the LDA document–topic distribution (θ_d). The resulting clusters were then labeled with predefined topic names according to their highest-probability assignments.

Flickr and YouTube corpora were modeled separately, each with its own LDA model. English texts were lemmatized, whereas Thai texts were tokenized without lemmatization. Cross-platform comparisons in the Discussion are therefore made at the level of interpreted topics rather than from a single joint LDA model.

3.2.2. YouTube Sentiment Analysis

YouTube comment sentiment was classified using a rule-based lexicon approach. This method was selected for three reasons specific to the context of this study. First, rule-based classifiers are fully auditable: every decision is traceable to explicit lexical rules, which is essential for spatial research where the interpretability of place-specific sentiment scores underpins geographic analysis and must be independently verifiable [45]. Second, the dataset consists of short comments (mean ≈ 14 tokens) containing highly domain-specific ecotourism vocabulary absent from general Thai pre-training corpora, limiting the effective advantage of transformer-based models such as WangchanBERTa [46,47]. Third, no annotated Thai ecotourism sentiment dataset exists for supervised fine-tuning; therefore, the use of a rule-based approach avoids the risk of poorly calibrated transfer learning.

The sentiment lexicon was constructed in two stages. An initial set was manually curated from Thai ecotourism and travel commentary. This set was subsequently expanded via corpus frequency analysis of all 8383 comments, identifying 1646 candidate terms (frequency ≥ 5), from which contextually relevant terms were selected. Three keyword categories were defined: positive, negative, and neutral. Positive keywords encompass expressions of appreciation, satisfaction, and perceived quality (e.g., 'beautiful', 'excellent', 'worthwhile', 'relaxing'), together with social-media colloquialisms (e.g., chill, wow). Negative keywords employ a three-tier weighting scheme: strong negatives (weight = 1.0) encode unambiguous dissatisfaction (e.g., 'bad', 'disappointed', 'dirty'), while weak negatives (weight = 0.5) describe physical or contextual conditions that frequently co-occur with positive ecotourism experiences (e.g., 'expensive', 'exhausted', 'difficult') and a single micro-negative term captures marginal dissatisfaction not covered by the other tiers. Neutral keywords comprise destination names, generic travel activity terms, and site-specific nouns (e.g., 'porter', 'cable car'). The final lexicon size was 68 keywords, comprising 29 positive, 14 negative (10 strong, 3 weak, and 1 micro-negative), and 25 neutral terms.

Sentiment classification was then performed at the token level using the three keyword categories, and sentiment scores were computed from effective keyword counts using three algorithmic refinements. First, a window-based negation rule was used to detect negation particles "not", "not really", and "did not" within a two-token window preceding any sentiment keyword, and the polarity of the sentiment keyword was accordingly reversed (e.g., "not beautiful" → negative, "not bad" → positive). This dynamic mechanism replaces the prior practice of enumerating compound negation forms as discrete static keywords, eliminating double-counting and generalizing negation handling to all keyword pairs. Second, weak negative terms were assigned half-weight (0.5) to reduce over-prediction of negative sentiment when such terms appeared in otherwise positive ecotourism contexts (e.g., 'exhausted but worth it'). Third, contrast detection was used to cancel the negative contribution of a keyword when a contrast particle occurred within six tokens after the negative keyword and a positive keyword followed (e.g., 'expensive but satisfying' is positive). Effective positive and negative scores were summed across all matched keywords. The label was determined by the higher score, and ties were resolved in favor of the positive class. Comments with zero net contributions in both categories were labeled neutral.

A random sample of 200 comments was manually annotated with sentiment labels (positive, neutral, and negative), and the results were compared with the automated predictions to assess the reliability of the sentiment classification. Classification performance was evaluated using overall accuracy, Cohen's κ coefficient, precision, recall, and F1-score [16].

3.3. Diversity Metrics

Flickr tags were analyzed to identify ecosystem and landscape attributes. The semantic diversity of user-generated content was quantified based on tag distributions across national parks. Diversity metrics commonly used in ecology and information theory were adopted for this analysis [18]. In this context, each unique tag was treated as an analogue of a landscape type, and its frequency of occurrence was treated as its abundance of that landscape. The Shannon diversity index [48] was computed using the vegan package [49] to compare ecology-related tag diversity and variation across national parks. The index was calculated using Equation (3).

$$H' = -\sum_{i=1}^{s} P_i \ln P_i \tag{3}$$

where H' is the Shannon–Weaver diversity index, which indicates landscape diversity; higher values represent greater heterogeneity. P_i is the proportion of individuals belonging to landscape type i relative to the total number of individuals in the reviews, and ln denotes the natural logarithm.

In addition, the tag entropy (D) was calculated using the same Shannon formulation, with an emphasis on its interpretation in information theory as a measure of unpredictability in tag usage, as shown in Equation (4). For example, if Flickr users at Doi Inthanon applied a wide range of tags with relatively balanced frequencies (e.g., "mountain," "waterfall," "birdwatching," and "view"), entropy would be high, reflecting diverse visitor perspectives. Conversely, if tags for the Similan Islands were dominated by a few recurring terms (e.g., "beach," "dive," and "reef"), entropy would be low, indicating a narrower discourse.

$$D = exp\ (H') = exp(-\sum_{i=1}^{s} P_i \ln P_i) \tag{4}$$

3.4. Topic–Sentiment Mapping

To examine the relationship between semantic content of visitor reviews and their emotional orientation, a topic–sentiment mapping was conducted. This approach captures both what visitors talk about (topics) and how they feel about those themes (sentiment), thereby highlighting thematic domains associated with positive, neutral, or negative perceptions [21]. Each comment was assigned a dominant topic label using LDA and a sentiment polarity using a rule-based lexicon method with negation handling. A cross-tabulation of topics and sentiment categories was then performed at the park level to quantify the relative distribution of sentiment across thematic domains. This enabled visualization of the associations between topics and sentiment orientations. Topics positioned closer to the positive category in an ordination space indicate stronger alignment with positive visitor discourse, whereas those located nearer the negative category reflect dissatisfaction.

3.5. Statistical Modeling

The statistical modeling framework was structured as a Diversity × Topic × Season matrix for sentiment. Flickr-derived landscape indicators were first aggregated by national park and season, whereas individual YouTube comments were used as the unit of analysis for logistic regression. Each comment was associated with a unique combination of national park and seasonal period. Seasons were categorized into three climatological regimes commonly recognized in Thailand: rainy (May–October), winter (November–February), and summer (March–April). Organizing the dataset in this way enabled the simultaneous examination of spatial heterogeneity across protected areas and temporal variability across seasons. This design is particularly suitable for ecotourism research, as patterns of visitor activity and perception fluctuate both geographically and seasonally [50]. The Diversity × Season × Topic matrix therefore provides a consistent analytical framework for integrating multi-source social media indicators, specifically Flickr-derived spatial–semantic features and YouTube-derived perceptual–sentiment features into a harmonized comparative analysis. Although the unit of analysis in the logistic regression is the individual YouTube comment, comments are nested within videos, parks, and seasons. In the present study, park–season-level predictors therefore enter the model as contextual variables, and higher-level random effects are not modeled explicitly.

To investigate the relationships between Flickr-derived spatial–semantic indicators and YouTube-derived perceptual outcomes, we employed a multi-stage statistical analysis. A binary logistic regression was used to examine how Flickr features predicted the likelihood of positive sentiment in YouTube comments [51–53]. Effective tag diversity ($exp(H')$), the relative frequency of landscape-related tags (e.g., "beach" and "forest"), and YouTube topic proportions for each season across all parks were included as predictors, while sentiment polarity (positive = 1, neutral and negative = 0) served as the dependent variable. The significance of model coefficients (β) was assessed using standard Wald tests to identify influential predictors. The model structure is expressed as:

$$logit\,\{P(POS{=}1)\} \sim \beta_0 + \beta_1(effective\ tag) + \beta_2(Flickr\ topic) + \beta_3(YouTube\ topic) + \beta_4(Season) \quad (5)$$

where β_0 is the intercept, and β_1–β_4 are coefficients representing the effects of Flickr- and YouTube-derived predictors, while the season factor captures temporal variability.

The model was fitted using a binomial logit link function; the reference categories were the rainy season, Flickr topic Other, and YouTube topic Feelings & Experiences. Regression coefficients were transformed into odds ratios (ORs) with 95% confidence intervals (CIs), and statistical significance was assessed using Wald *z*-tests. Model performance was evaluated using the null deviance, residual deviance, Akaike Information Criterion (AIC), and McFadden's pseudo-R^2. Multicollinearity among predictor variables was assessed using the generalized variance inflation factor (GVIF), with adjusted GVIF values ($GVIF^{(1/(2Df))}$) reported for categorical predictors [54]. Thus, the magnitude and direction of the associations between landscape-related semantic indicators and the probability of expressing positive sentiment across parks and seasons were quantified.

The influence of the COVID-19 pandemic on the observed relationships was assessed via a sensitivity analysis by adding a binary pandemic-period indicator to the pooled logistic regression model (2020–2021 = 1; other years = 0).

3.6. Spatial Analysis Procedure

Spatial patterns of social sensing indicators were analyzed to reveal the geographic distribution of thematic topics, sentiment polarity, and tag diversity across the five national parks. Geotagged Flickr photographs were used to calculate landscape dominance and spatial diversity at the park level. For YouTube, topic categories and net sentiment polarity were derived from video metadata, with net sentiment calculated as the difference between positive and negative proportions. Flickr thematic patterns were visualized using dot maps, whereas YouTube topics were represented with pie charts, where slice sizes indicated topic proportions and pie sizes corresponded to net sentiment magnitude.

A net sentiment intensity index was calculated to summarize the overall emotional tendency for each national park:

$$\text{Net Sentiment} = \frac{N_{positive} - N_{negative}}{N_{positive} + N_{Neutral} + N_{negative}} \quad (6)$$

where $N_{positive}$, $N_{Neutral}$, and $N_{negative}$ denote the numbers of comments classified as positive, neutral, and negative, respectively. The net sentiment intensity index ranges from −1 to +1, where positive values indicate predominantly positive visitor sentiment, negative values indicate predominantly negative sentiment, and values close to zero indicate balanced emotional responses. The radius of each pie chart is proportional to the net sentiment intensity.

3.7. Expert Evaluation

The face validity of the identified topics and park-level sentiment profiles was assessed via an expert validation exercise in accordance with established expert elicitation guidelines [55]. Four experts with direct professional experience in at least one of the five parks were purposively selected. The experts included an ecotourism researcher, a national park ranger, a park manager, and a tourism practitioner. Experts evaluated (1) the thematic coherence of the three LDA-derived topics based on their ten most representative keywords and (2) the consistency of park-level sentiment distributions with their professional observations using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Inter-rater agreement was assessed using Fleiss' kappa (κ) [56] and interpreted according to the criteria described in a previous study [57]. A topic or sentiment profile was considered to demonstrate acceptable face validity if the mean expert rating was ≥ 3.5 and at least 70% of ratings were ≥ 4.

4. Results

4.1. Descriptive Statistics and Text Analysis

The annual distribution of Flickr and YouTube posts between 2018 and 2024 is presented in Figure 3. Flickr activity peaked in 2019 (>23,000 posts) but declined sharply during the COVID-19 period (2020–2021) and partially recovered after 2022. In contrast, YouTube posting patterns showed substantially less disruption, remaining relatively stable and peaking in 2021. This discrepancy highlights platform-specific demographic biases: Flickr's decline mirrors the collapse of international tourism, as the platform is predominantly used by global visitors, whereas YouTube appears less sensitive to mobility restrictions, indicating a strong domestic user base that continued to engage in "virtual tourism" and digital storytelling even during periods of reduced physical travel. These differences underscore the importance of integrating multi-platform data to mitigate representation bias associated with any single social media source.

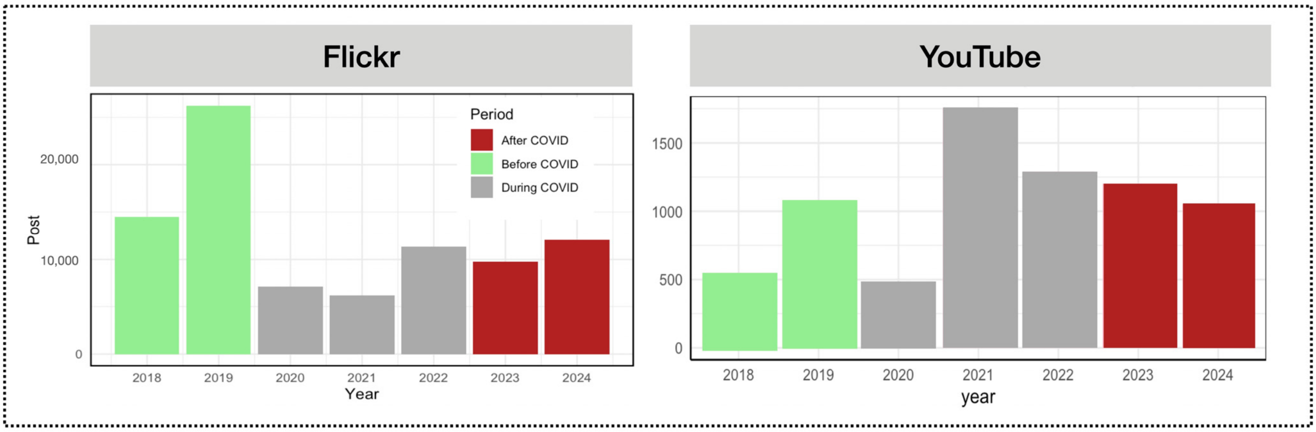

Figure 3. Annual number of Flickr posts (**left**) and Thai YouTube posts (**right**) from 2018 to 2024.

Building on these platform-wide patterns, the selected analytical corpus comprised 1715 geotagged Flickr photographs with valid tags and 8383 cleaned YouTube comments across the five national parks. Flickr photographs were retained after applying spatial, temporal, and metadata filters, and non-informative tags were removed, yielding a tag vocabulary that was used in subsequent diversity and topic modeling analyses. The YouTube dataset, comprising comments on park-related videos, was cleaned to remove URLs, markup, and emojis and prepared for bilingual NLP. Spatial coverage was ensured by including both forest–mountain and marine parks. Comments in the YouTube corpus were distributed across parks as follows: Phu Kradueng (3022 comments), Koh Chang (2196), Khao Yai (2040), Doi Inthanon (715), and Similan (410), reflecting a mix of upland and coastal ecotourism destinations. Comments were distributed across seasons as follows: winter (3528), rainy (3496), and summer (1359), corresponding to Thailand's three climatological regimes and providing sufficient data for the Diversity × Season × Topic analyses reported in Sections 4.2–4.4.

Figure 4 shows the 30 most frequently occurring Flickr tags. Word size reflects tag frequency, emphasizing the key themes associated with each park. Similan is dominated by marine recreation tags ("scuba," "diving," "underwater"), while Doi Inthanon is characterized by biodiversity-related terms (bird- and wildlife-related terms). Khao Yai features scenic keywords ("species," "birds," "water"), consistent with its forested landscape. Koh Chang emphasizes marine and travel-related themes ("sea," "beach," "ferry"), while Phu Kradueng highlights trekking and mountainous terrain ("steps," "viewpoint," "steep"). These differences in keyword prominence illustrate how visitors perceive and represent each park, aligning closely with their distinct ecological and recreational profiles.

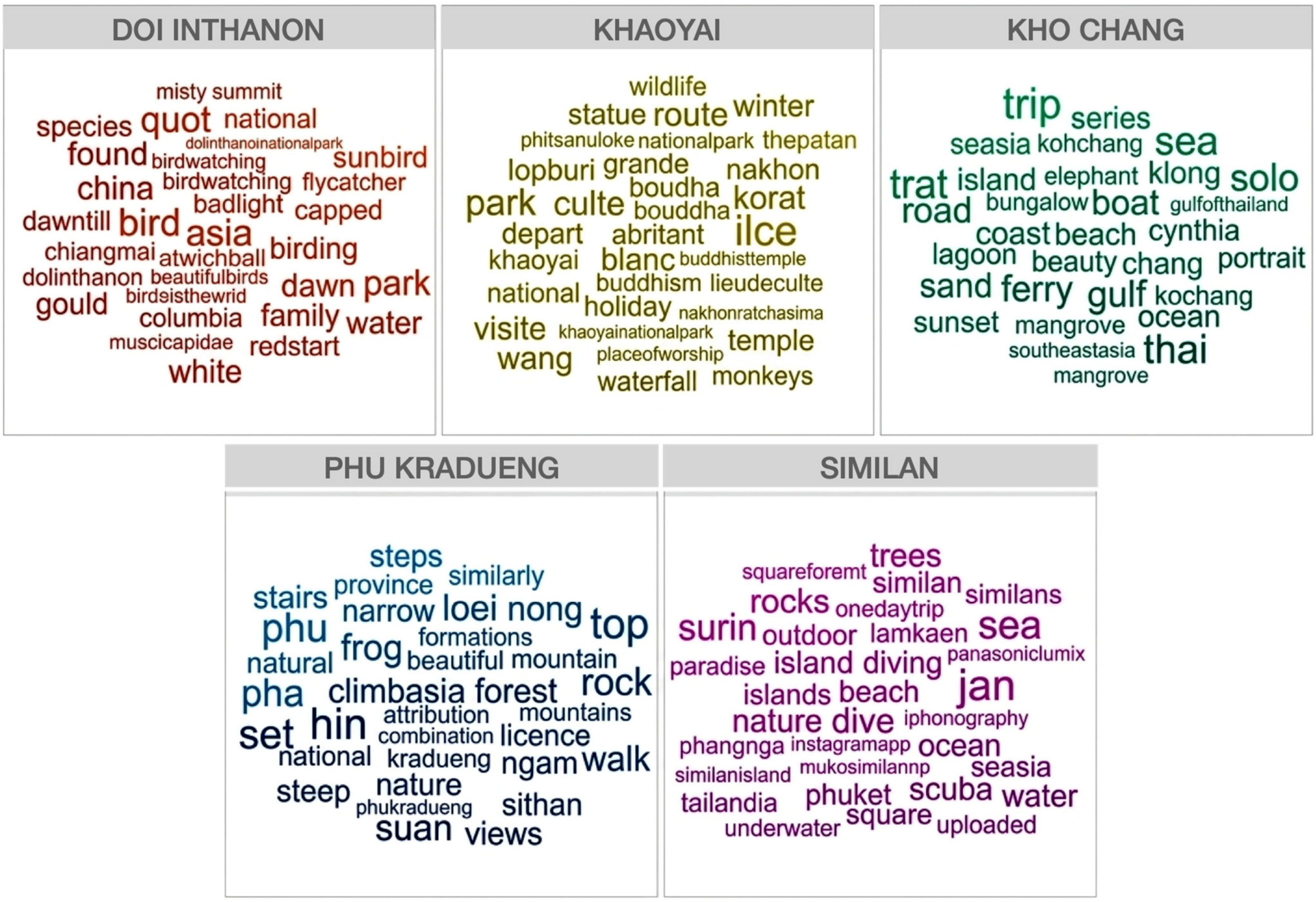

Figure 4. Word cloud of the top thirty most frequently occurring Flickr tags for each national park.

4.2. Topic–Sentiment Analysis

For Flickr, candidate topic models with k = 3, 4, 5, 6, 8, and 10 were evaluated using topic coherence, topic diversity, perplexity, and manual assessment of semantic interpretability. Although models with larger numbers of topics produced lower perplexity values, perplexity was not used as the primary criterion for model selection because it generally decreases as the number of topics increases. Instead, greater emphasis was placed on topic coherence and semantic interpretability. Manual inspection indicated that models with higher numbers of topics tended to divide conceptually related themes into more specific subtopics, whereas the three-topic model provided broader, clearly distinguishable, and semantically meaningful themes that were more appropriate for comparing visitors' perceptions across protected areas. Accordingly, the Flickr topics were categorized into three semantic dimensions: Topic 1 (Marine), Topic 2 (Forest & Mountain), and Topic 3 (National Park Landscapes).

For YouTube, the three-topic model was selected because it achieved the highest coherence score (0.504) among the evaluated models while producing coherent and interpretable thematic structures. Although this coherence value is moderate, such values are commonly reported for short social media texts because of lexical sparsity, informal language, and high variability in user-generated content. Topic diversity was also maximized (1.0), indicating that the extracted topics remained distinct. Manual examination of the representative keywords and associated comments further confirmed that the three-topic solution effectively captured three meaningful dimensions of visitors' perceptions: Topic 1 (Travel & Activities), Topic 2 (Nature & Landscapes), and Topic 3 (Feelings & Experiences) (Tables 1 and 2).

Table 1. LDA model evaluation for YouTube and Flickr datasets based on coherence, perplexity, and topic diversity.

YouTube				Flickr			
k	**Coherence**	**Perplexity**	**Diversity**	**k**	**Coherence**	**Perplexity**	**Diversity**
3	0.5041	1120.6142	1.0000	3	0.7187	257.6674	0.9000
6	0.4900	1773.6853	0.9333	10	0.6225	223.0447	0.7700
4	0.4713	1331.2032	0.925	4	0.6129	254.2300	0.8500
5	0.4676	1568.3966	0.8800	5	0.5891	243.8118	0.8200
8	0.4533	2244.1196	0.9000	6	0.5313	234.5636	0.8333
10	0.4145	2755.9457	0.8700	8	0.5244	223.3840	0.7625

The semantic meaning of each topic was determined through manual interpretation of the highest-probability keywords together with representative social media content. For Flickr, Topic 1 (Marine) represents coastal tourism and marine recreation, Topic 2 (Forest & Mountain) reflects biodiversity-related activities such as wildlife observation, and Topic 3 (National Park Landscapes) captures iconic scenic environments. In contrast, the YouTube topics emphasize experiential and affective dimensions of tourism. Topic 1 (Travel & Activities) represents recreational practices and travel experiences, Topic 2 (Nature & Landscapes) highlights environmental features, and Topic 3 (Feelings & Experiences) reflects visitors' emotional responses. These findings indicate that image-based platforms primarily capture visual landscape characteristics, whereas video-sharing platforms reveal richer narrative and emotional engagement with protected areas.

Table 2. Top representative keywords for each LDA topic for Flickr and YouTube (k = 3).

Topic (YouTube)	**Top Keywords**
Topic1: Travel and Activities	transportation, hiking, diving, camping, guides, park services, logistics
Topic 2: Nature and Landscapes	forests, waterfalls, sea, coral reefs, mountains, scenic views
Topic 3: Feelings and Experiences	happiness, excitement, challenge, friendship, romance, memorable moments
Topic (Flickr)	Top Keywords
Topic1: Marine (coastal tourism)	island, sea, beach, coastal, marine, shoreline, coastal landscape
Topic 2: Forest & Mountain (birdwatching and wildlife observation)	bird, white-capped redstart, birdwatching, birding, wildlife, dawn, biodiversity
Topic 3: Others (National Park landscapes)	national park, protected area, landscape, forest, nature, scenic environment

A lexicon-based sentiment analysis was applied to YouTube comments to further examine emotional orientations within these thematic domains, and the results are shown in Figure 5. Overall, the neutral sentiment was dominant across all five destinations, accounting for 45–62% of the comments. These results reflect the prevalence of descriptive and activity-oriented content and indicate that a substantial portion of commentary documents rather than evaluates the experience.

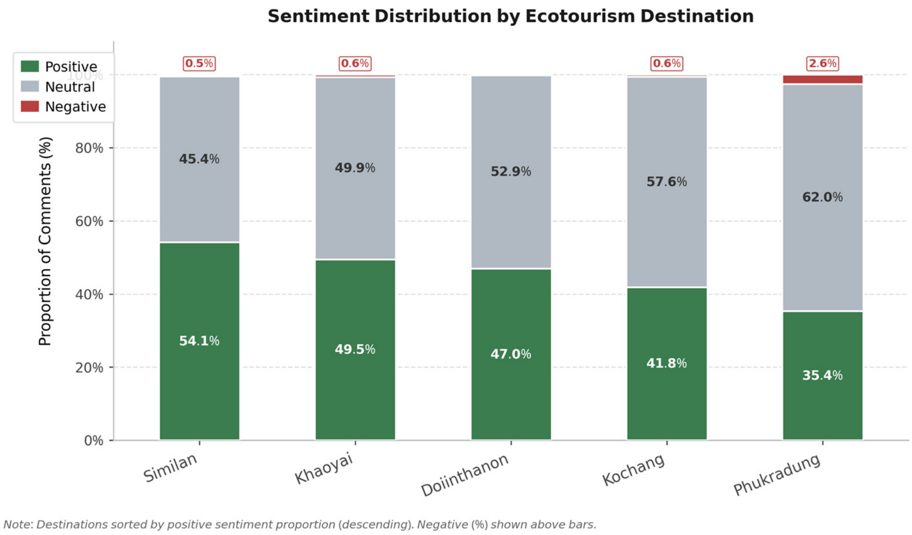

Figure 5. Distribution of sentiment polarity (positive, neutral, negative) in YouTube comments.

The Similan Islands recorded the highest proportion of positive sentiment (54.1%), followed by Khao Yai (49.5%) and Doi Inthanon (47.0%), suggesting that destinations offering high scenic quality and well-managed visitor experiences elicit stronger affective responses. Negative sentiment was the highest at Phukradung (2.6%), while the other four destinations recorded low proportions of negative sentiments (≤0.6%). Word frequency analysis of the negative comments revealed that the most prominent terms at Phukradung were difficult, exhausted, porter, and cable car, indicating that the negative discourse centers primarily on physical accessibility challenges and concerns over porter services following the suspension of the cable car, rather than dissatisfaction with the destination itself. The terms sympathy and age further suggest that visitor concerns extend to the suitability of the trail for elderly visitors and young children.

Records with missing labels were excluded, and thus, 190 comments were retained for validation. The automated predictions were compared with the manual annotations using standard classification metrics. The validation results are summarized in Table 3. The classifier achieved an overall accuracy of 79.5%, a weighted F1-score of 0.798, and a Cohen's kappa coefficient of 0.690, indicating substantial agreement between manual and automated sentiment classification. Performance was the highest for the positive class (F1 = 0.828), whereas the neutral and negative classes achieved F1-scores of 0.788 and 0.755, respectively. Most classification discrepancies occurred between positive and neutral comments, reflecting the subjective nature of mildly positive expressions common in social media comments.

Table 3. Validation results of the automated sentiment classification against manually annotated comments.

Sentiment	Precision	Recall	F1-Score
Positive	0.938	0.741	0.828
Neutral	0.806	0.771	0.788
Negative	0.627	0.949	0.755
Overall accuracy	79.5%		
Cohen's κ	0.690		
Weighted F1-score	0.798		

4.3. Tag Diversity

Seasonal variation in effective tag diversity (exp H') is illustrated in Figure 6. This metric serves as a quantitative indicator, reflecting the variety of landscape attributes and activities captured by visitors across the five national parks. Koh Chang and Doi Inthanon exhibit the highest levels of tag diversity, particularly during the summer and winter seasons, reflecting their heterogeneous landscapes and diverse activity profiles. Khao Yai shows moderate diversity year-round, consistent with its mixed forest and wildlife environments. Seasonally, summer and winter generally display higher tag diversity than the rainy season, likely reflecting peak visitation periods and more favorable weather conditions for exploring varied landscapes. In contrast, Similan and Phu Kradueng consistently show lower tag diversity, suggesting more homogeneous landscape features or limited seasonal variation in visitor activities. Overall, these patterns highlight both spatial and temporal heterogeneity of tourism. While social media data is subject to data density biases, capturing variations across seasons and park types provides an interpretable indicator of how landscape variety corresponds to cognitive and behavioral complexity.

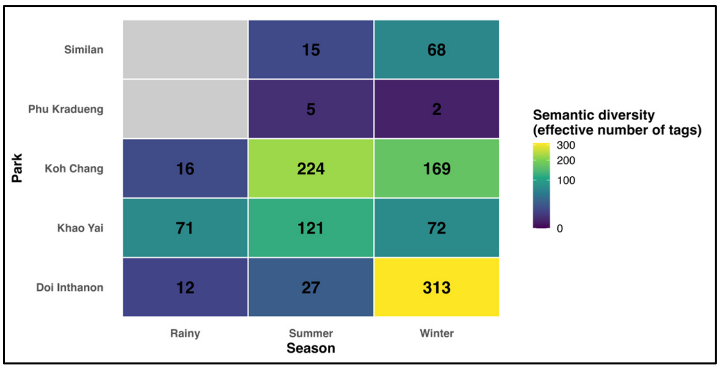

Figure 6. Seasonal variation in semantic diversity, expressed as the effective number of tags (exp(H')), calculated from the Shannon diversity index of Flickr tags across five national parks in Thailand. Higher values indicate greater semantic diversity in visitor-generated tags. Grey cells indicate park–season combinations with no available geotagged Flickr tags; therefore, semantic

https://doi.org/10.3390/technologies14080456

diversity could not be calculated. Note: Diversity estimates for Park × Season combinations with fewer than 10 Flickr tags should be interpreted with caution because of increased sampling variability.

For park–season combinations with very small tag counts (e.g., $n < 10$), the resulting diversity estimates are less reliable because of increased sampling variability. These observations are retained to preserve the complete seasonal representation of the dataset but should be regarded as exploratory rather than used for strong comparative inference.

4.4. Statistical Modeling of Perceptual Outcomes

Building on these spatio-semantic indicators, we examined the relationship between the indicators and perceptual outcomes derived from YouTube data using a binary logistic regression model, operationalized as a Diversity × Topic × Season framework for positive sentiment. The unit of analysis was the individual YouTube comment ($n = 8383$), with Shannon diversity, Flickr-derived topic, season, and dominant YouTube LDA topic used as the predictors of positive sentiment. Regression coefficients were reported as OR with 95% CI. Model diagnostics, including McFadden's pseudo-R^2, AIC, residual degrees of freedom, and GVIF, were used to assess model fit and multicollinearity. Multicollinearity diagnostics indicated no evidence of problematic collinearity among the predictors, with adjusted GVIF values ($GVIF^{(1/(2Df))}$) ranging from 1.060 to 1.623, well below commonly accepted thresholds of 5. The regression coefficients are shown in Figure 7, and the model diagnostics are summarized in Table 4.

Table 4. Summary statistics and diagnostics of the logistic regression model.

Statistic	Value
Number of observations	8383
Null degrees of freedom	8382
Residual degrees of freedom	8375
Null deviance	11,428
Residual deviance	11,243
Akaike Information Criterion (AIC)	11,259
McFadden's pseudo-R^2	0.0161
Maximum Variance Inflation Factor (VIF)	1.623

Note: The regression model was fitted using individual YouTube comments as the unit of analysis. Model fit is summarized using McFadden's pseudo-R^2. Multicollinearity was assessed using the generalized variance inflation factor (GVIF); the adjusted $GVIF^{(1/(2 \times Df))}$ values ranged from 1.060 to 1.623, indicating negligible multicollinearity among predictors.

This modeling framework enables an assessment of how landscape-related spatio-semantic indicators correspond to public emotional responses across thematic domains (Figure 7).

Among continuous predictors, Shannon tag diversity showed a significant positive association with positive sentiment (OR = 1.19, 95% CI: 1.09–1.31, $p < 0.001$), indicating that the higher semantic diversity of Flickr tags was associated with a greater likelihood of positive sentiment.

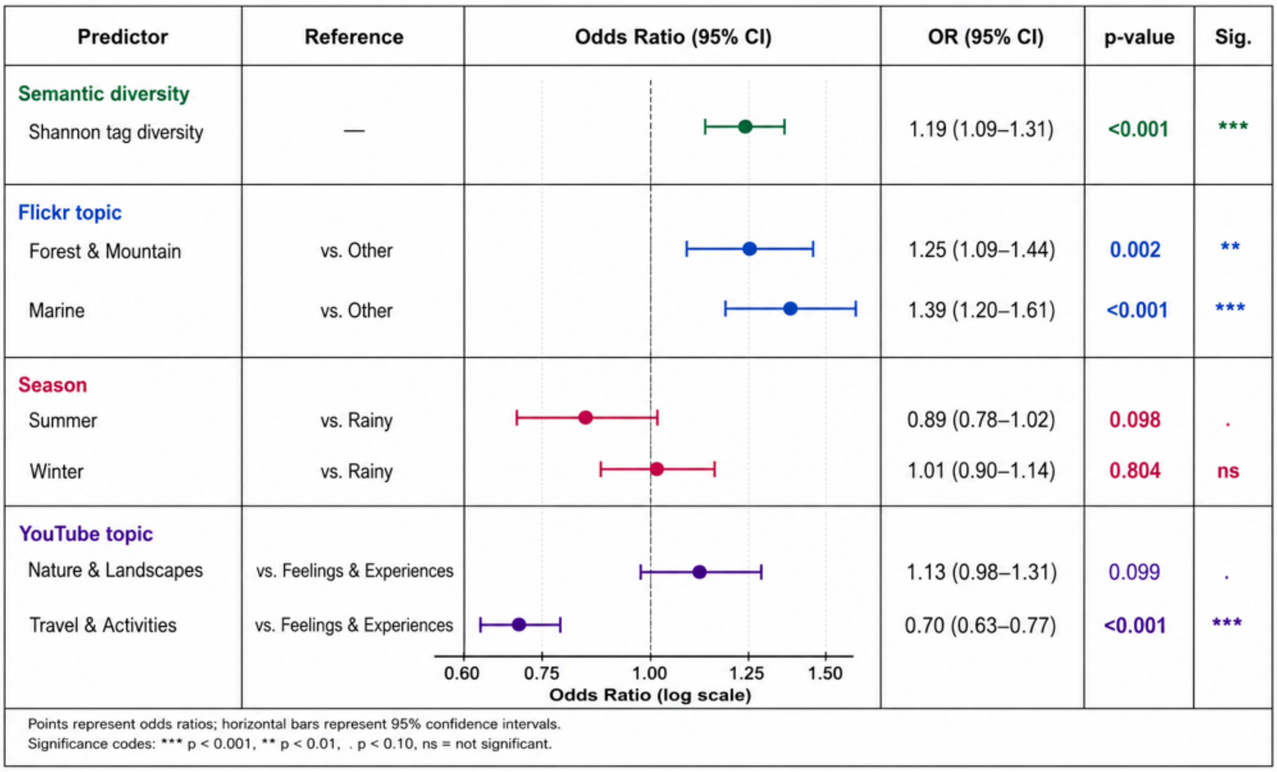

Figure 7. Odds ratios (log scale) from the binary logistic regression model illustrating the association between Flickr-derived landscape topics, seasonal variables, and YouTube thematic topics and positive sentiment. Points represent estimated ORs, horizontal bars indicate 95% confidence intervals (CIs), and the dashed vertical line denotes the null reference (OR = 1). Colors indicate predictor categories: green = semantic diversity, blue = Flickr topics, pink = season, and purple = YouTube topics.

Forest & Mountain tags (OR = 1.25, 95% CI: 1.09–1.44, $p = 0.002$) and Marine (OR = 1.39, 95% CI: 1.20–1.61, $p < 0.001$) derived from Flickr were associated with significantly higher odds of positive sentiment relative to the reference category (Other).

Seasonal effects were comparatively weak. Relative to the rainy season, the summer season showed a tendency toward lower odds of positive sentiment (OR = 0.89, 95% CI: 0.78–1.02, $p = 0.098$), although this effect was only marginally significant. Statistically significant differences were not observed between the winter and rainy seasons (OR = 1.01, 95% CI: 0.90–1.14, $p = 0.804$).

Among the YouTube-derived thematic variables, Nature & Landscapes did not differ significantly from the reference category (Feelings & Experiences) (OR = 1.13, 95% CI: 0.98–1.31, $p = 0.099$). In contrast, Travel & Activities was associated with substantially lower odds of positive sentiment (OR = 0.70, 95% CI: 0.63–0.77, $p < 0.001$), indicating that comments focused primarily on travel logistics, transportation, hiking, camping, diving, guides, fees, or related activities were less likely to express positive sentiment compared to those emphasizing personal feelings and experiences.

A COVID-adjusted model including a binary indicator for the pandemic period (2020–2021) was also fitted for sensitivity analysis. The estimated OR remained highly similar to those of the primary model, and McFadden's pseudo-R^2 changed only marginally (0.01613 in the primary model vs. 0.01617 in the COVID-adjusted model), suggesting that the principal associations were not materially influenced by pandemic-period observations.

4.5. Spatial Analysis

The spatial distributions of Flickr photographs and YouTube-derived topics reveal distinct thematic patterns across the five national parks (Figure 8). Flickr photographs reflect the dominant ecological characteristics of each park, with marine-related photographs concentrated in Koh Chang and Similan, whereas forest and mountain photographs predominate in Doi Inthanon, Khao Yai, and Phu Kradueng. Despite these differences, the Flickr Shannon diversity index (H′) remained relatively high across parks, ranging from 4.47 in Phu Kradueng to 5.94 in Khao Yai, indicating diverse visitor-generated photographic content.

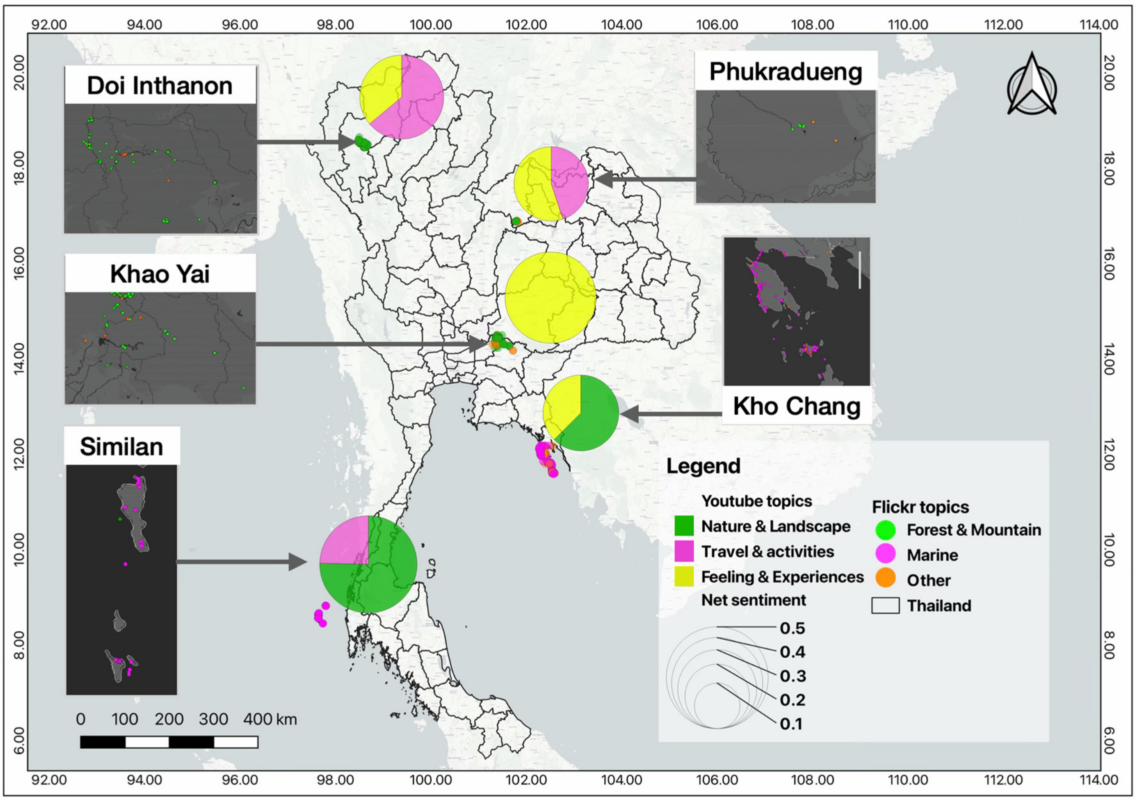

Figure 8. Spatial distribution of social sensing indicators across the five national parks. Colored Flickr points represent the dominant thematic categories of geotagged photographs: Forest & Mountain (green), Marine (magenta), and Other (orange). Pie charts illustrate the proportional distribution of YouTube-derived themes, including Nature & Landscape (green), Travel & Activities (magenta), and Feelings & Experiences (yellow). Pie size is proportional to the magnitude of the net sentiment index. Insets highlight the spatial distribution of Flickr photographs within each national park.

The YouTube topic distributions also differed markedly among parks. Doi Inthanon exhibited a relatively balanced distribution between Travel & Activities (63.9%) and Feelings & Experiences (36.1%), whereas Phu Kradueng showed a similar pattern, with Feelings & Experiences (55.1%) slightly exceeding Travel & Activities (44.9%). Khao Yai was entirely dominated by Feelings & Experiences (100%), suggesting that visitors primarily emphasized emotional responses and overall travel experiences. In contrast, Nature & Landscape represented the largest topic in Koh Chang (62.5%) and Similan (75.4%), with the remaining content attributed to Feelings & Experiences (37.5%) in Koh Chang and Travel & Activities (24.6%) in Similan.

Variation in pie-chart size reflects differences in net sentiment intensity across parks. Similan exhibited the highest net sentiment (0.456), followed by Khao Yai (0.402), Doi

Inthanon (0.340), Koh Chang (0.277), and Phu Kradueng (0.267). Overall, the spatial patterns indicate that ecosystem characteristics influence not only the spatial distribution of visitor-generated content but also the thematic emphasis and emotional expression captured from social sensing data.

4.6. Expert Validation

Four experts—an ecotourism researcher, a national park ranger, a park manager, and a tourism practitioner—completed an elicitation exercise. The mean (M) face validity scores ranged from 3.9 to 4.3 across the three LDA-derived topics (Nature & Landscapes: M = 4.3, standard deviation (SD) = 0.6; Travel & Activities: M = 4.1, SD = 0.7; Feelings & Experiences: M = 3.9, SD = 0.8), with 78% of experts rating all topics as adequate or highly valid (score ≥ 4). Fleiss' kappa indicated substantial inter-rater agreement (κ = 0.67), consistent with the Landis–Koch benchmark for substantial agreement (0.61–0.80). Moreover, 87% of the experts indicated that the per-park sentiment profile distributions were consistent with their observations.

5. Discussion

5.1. The Diversity × Topic × Season Framework

In this study, the spatio-semantic dimensions of tourist perceptions across five national parks in Thailand were examined by integrating geotagged Flickr imagery with YouTube narratives. Building on this multi-platform dataset, we developed a Diversity × Topic × Season framework that links spatial, temporal, and thematic variations in visitor expressions. In this framework, geotagged photographs provide spatial anchors for landscape attributes, whereas topic modeling and sentiment analysis applied to textual data reveal visitors' descriptions and emotional evaluations of those environments. A central feature of this framework is the use of effective tag diversity $exp(H')$ as a conceptual and methodological bridge between ecological metrics and semantic diversity. This bridge relies on the shared use of Shannon entropy, a diversity metric originally formulated to quantify habitat heterogeneity, operationalized here as an analog for linguistic richness in visitor narratives. We highlight that this bridge is conceptual rather than empirical. In this study, we do not directly test correlations between $exp(H')$ and independent ecological indicators, such as species richness, remote-sensing-derived habitat heterogeneity, or structural complexity measured using LiDAR. Instead, we use the landscape type (forest–mountain vs. marine) as a categorical proxy for ecological heterogeneity, consistent with prior work linking habitat diversity to diversity in visitor experience. A major strength of the developed framework lies in its use of diversity-oriented NLP metrics that translate linguistic richness into spatially interpretable indicators. By mapping topic prevalence, sentiment polarity, and tag diversity and seasons, the framework enables interpretable links between ecological patterns and semantic expression and provides a transferable template for integrating social sensing data into protected-area management and ecotourism planning.

5.2. Landscape Heterogeneity and Semantic Diversity

Consistent differences emerged between forest–mountain and marine landscapes. Forest parks (Doi Inthanon, Khao Yai) exhibited greater semantic diversity (exp(H')) because visitors engage with multiple components such as wildlife and trails. This pattern is in good agreement with information processing theory, which suggests that environments characterized by complexity and mystery stimulate greater cognitive engagement, producing richer and more varied linguistic outputs. In contrast, marine parks (Similan, Koh Chang) showed narrower diversity, but stronger positive sentiment tied to diving. This relationship between landscape heterogeneity and textual richness underscores the role of environmental variety in shaping tourists' linguistic expression and emotional tone, analogous to how species diversity supports ecological complexity in natural systems [34].

Spatially, the distribution of thematic and sentiment indicators reveals how landscape context shapes perception. From a landscape-aesthetics perspective [10–12], visitor tags at forest parks, such as "waterfall," "birdwatching," and "viewpoint," reflect aesthetic appreciation of the layered, multi-textured environments, whereas marine park tags cluster around a single dominant aesthetic (underwater visibility). This contrast demonstrates the translation of aesthetic complexity into semantic complexity, supporting the operationalization of Shannon entropy as a measurable index of perceived landscape richness. Marine environments elicited concise, affect-rich vocabularies tied to visual clarity and recreational value, whereas forest ecosystems generated multidimensional narratives that encompass both ecological appreciation and personal experience. Temporally, summer and winter seasons showed the highest diversity and positivity, coinciding with favorable environmental conditions and peak tourism periods. This pattern suggests that climatic comfort and landscape visibility directly enhance affective language, linking environmental quality to positive tourist sentiment.

These perceptual patterns also reflect underlying preferences: international tourists predominantly shape coastal imagery on Flickr, whereas domestic visitors generate more narrative-rich YouTube content from forested parks. This interplay between landscape context and user demography not only defines the cultural ecology of ecotourism but also shapes distinct pathways through which different groups develop environmental awareness [58,59]. From a management perspective, these findings suggest that textual indicators can serve as early signals of visitor satisfaction and thematic emphasis. These insights can support tourism promotion strategies by highlighting landscape elements and experiences that visitors frequently appreciate, enabling destination managers to emphasize these features in communication and marketing materials. For example, higher semantic diversity in forest parks such as Doi Inthanon and Khao Yai suggests diverse visitor engagement with ecological features, including camping activities and wildlife observation (e.g., elephants and birds). In contrast, marine parks such as Similan and Koh Chang may benefit from targeted tourism promotion focusing on diving and scuba activities. Spatial patterns of semantic diversity and topic concentration may help identify areas associated with distinct visitor perceptions. Such information can guide adaptive zoning and visitor management, for example, by distinguishing zones [60] from areas where activity-based experiences are more prevalent, while maintaining conservation priorities in protected landscapes [61,62].

Our framework aligns with Thailand's Five-Year Plan (2023–2027) and links social sensing to Blue Carbon and REDD+ initiatives [3,63]. Locations with low diversity and concentrated topics may benefit from interpretive interventions designed to broaden engagement beyond a few iconic features. Encouraging community-based participation in digital data collection, content creation, and perception analysis can enhance inclusivity and strengthen shared ownership of conservation outcomes that reflect local knowledge, cultures, values, and lived experiences. Integrating social-sensing and citizen science insights into national Blue Carbon and REDD+ strategies transforms online perception into a continuous feedback loop that strengthens adaptive management and fosters sustainable tourism across Thailand's land–sea interface.

From a theoretical perspective, the proposed Diversity × Topic × Season framework extends existing work on environmental perception and tourism experience by conceptualizing semantic diversity as an indicator of visitor perception. Whereas conventional sentiment analysis primarily captures the emotional valence of visitor

https://doi.org/10.3390/technologies14080456

responses, semantic diversity reflects the breadth of concepts through which visitors perceive, interpret, and describe natural environments. High semantic diversity therefore suggests that visitors engage with protected areas through multiple environmental meanings, recreational experiences, and conservation-related values rather than through a single dominant perception. In this sense, semantic diversity is not merely a descriptive linguistic statistic but a conceptual proxy for the diversity of environmental meanings that underpin ecotourism perceptions in protected-area settings.

5.3. Visitor Sentiment and Statistical Modeling

The LDA analysis revealed platform-specific thematic patterns. Flickr topics primarily represent landscape types and tourism destinations, whereas YouTube topics emphasize visitor experiences and emotional responses. The positive association between semantic diversity and positive sentiment suggests that linguistic variety can serve as an indicator of visitor satisfaction and perceptual engagement. YouTube narratives showed a predominance of neutral tones, indicating that many comments are descriptive in nature, often referring to locations, activities, or travel experiences rather than expressing strong emotional reactions.

Nevertheless, the concentration of positive sentiment around the Nature & Landscapes topic supports the hypothesis that natural scenes evoke calm, restorative, and awe-related responses. This finding is consistent with environmental perception theory, which posits that natural settings reduce cognitive fatigue and promote psychological well-being, a mechanism central to restorative environments research. The higher positivity observed for the forest and mountain contexts reflects how immersion in biodiverse landscapes fulfills visitors' need for fascination and escape, which are core components of Attention Restoration Theory. The lower positivity observed in the Travel & Activities topic may reflect the logistical or experiential challenges of visiting remote protected areas, such as accessibility constraints, crowding, or weather-related disruptions [27,64].

The manual validation results support the reliability of these sentiment patterns. Although the overall agreement between automated and manual sentiment classification was moderate (Cohen's $\kappa = 0.69$), disagreements were observed primarily between the positive and neutral categories. Examination of the manually validated comments indicated that many disagreements involved brief expressions of desire, appreciation, nostalgia, or mixed sentiment. For example, comments such as "Being a porter is difficult, but they're amazing" contain both negative contextual information ("difficult") and positive evaluative sentiment ("amazing"). Although the overall sentiment expressed by the user is positive, automated sentiment analysis models may misclassify such comments because they contain conflicting lexical cues and require contextual interpretation beyond individual sentiment-bearing words. This finding is consistent with previous studies showing that automated sentiment analysis of implicit sentiment and context-dependent expressions in user-generated content is challenging [65]. Therefore, the sentiment analysis results presented in this study should be interpreted as robust indicators of overall visitor perception patterns rather than exact classifications of individual comments.

Although several predictors in the logistic regression were statistically significant, the very low McFadden's pseudo-R^2 (0.0161) indicates that the model explains only a small proportion of the variation in visitor sentiment. Given the large sample size (n = 8383), statistically significant odds ratios must therefore be interpreted with caution, as they represent relatively weak practical effects rather than strong predictive relationships. Visitor sentiment is likely influenced by numerous unmeasured factors, including individual preferences, trip characteristics, environmental conditions, and contextual

https://doi.org/10.3390/technologies14080456

aspects of user-generated content. Consequently, the present analysis is best viewed as identifying statistically detectable correlates of visitor sentiment rather than providing a comprehensive explanatory model.

5.4. Limitations and Future Directions

Despite its contributions, this study has certain limitations. The findings are based exclusively on active social media users, who may not be representative of all park visitors. Social sensing data from Flickr and YouTube further depend on platform-specific demographics, with Flickr attracting photography-oriented international users and YouTube primarily reflecting domestic users interested in informational content. User-generated content is also unevenly distributed across parks and seasons, creating data density biases. In addition, sparse Flickr activity in some park–season combinations limits the robustness of the corresponding entropy estimates. The main patterns reported here therefore reflect park–season contexts with sufficiently large tag counts, and diversity values for very low-count cells should be treated as indicative rather than definitive.

Moreover, subtle or mixed emotions may be underrepresented in Thai sentiment analysis owing to linguistic nuances. In addition, topic modeling and sentiment classification depend on algorithmic assumptions that may limit thematic precision. These factors should be considered when generalizing the findings.

In terms of the statistical design, Flickr-derived semantic indicators were aggregated at the Park × Season level and embedded within a Diversity × Topic × Season matrix as contextual predictors of individual comment sentiment, which entails an ecological inference risk and likely contributes to the low McFadden's pseudo-R^2. The Diversity × Topic × Season framework therefore captures contextual associations rather than individual-level causal mechanisms, reinforcing the need for cautious interpretation of the regression results.

In addition, the logistic regression model did not explicitly account for the hierarchical structure of the data, as individual YouTube comments were nested within videos, parks, and seasons. Consequently, the assumption of complete independence among observations may not have been fully satisfied. This limitation, together with the very low McFadden's pseudo-R^2, reinforces our interpretation of the regression results as exploratory contextual associations rather than strong predictive effects. Future studies could employ mixed-effects logistic regression to explicitly model these hierarchical dependencies.

Recent multilingual transformer-based models have demonstrated superior performance in many sentiment analysis tasks, particularly in capturing contextual and implicit sentiment. Although the rule-based approach demonstrated moderate reliability for the purposes of this study, it may be less effective than domain-adapted transformer-based language models in capturing implicit sentiment, sarcasm, and context-dependent expressions. Thai transformer models (e.g., WangchanBERTa) fine-tuned on an annotated ecotourism corpus may be evaluated in future work, and their performance may be compared with that obtained via the rule-based approach with the same dataset. Such comparisons can help determine whether the additional model complexity and computational costs translate into meaningful improvements in understanding visitor perceptions.

To partially address the generalizability concern, a structured expert elicitation exercise was conducted, in which ecotourism researchers, national park rangers, and tourism practitioners assessed the face validity of the identified topics and the consistency of per-park sentiment profiles. Integrating multi-lingual, multi-platform data further helps mitigate platform-specific biases and yields a more multi-dimensional understanding of ecotourism perception, capturing visual engagement, linguistic

https://doi.org/10.3390/technologies14080456

interpretation, emotional tone, and temporal variations in how visitors experience protected areas. Although the expert elicitation provided useful face validation for the identified topics and sentiment profiles, the panel comprised only four experts. Consequently, the findings should be interpreted as qualitative supporting evidence rather than formal external validation or statistically generalizable conclusions. Future studies should include larger and more diverse expert panels to further strengthen the external validity of the evaluation.

Nevertheless, future research should triangulate these results against independent visitor surveys and extend the qualitative dimension through longitudinal structured interviews or multi-round expert consultation (e.g., Delphi technique) to further assess the external validity of social sensing indicators across different protected-area contexts.

6. Conclusions

This study developed an NLP-based framework for mapping how landscapes of Thailand's protected areas shape visitor perceptions by leveraging social sensing data from Flickr and YouTube. By integrating geographic patterns with emotional and visual responses, we demonstrated that ecosystem type is a primary determinant of both thematic content and visitor sentiment. Our findings contribute to the Diversity × Topic × Season matrix as a reproducible workflow that integrates topic modeling and Thai sentiment analysis. This advances geospatial NLP by offering a conceptual and methodological bridge that translates abstract social media patterns into actionable spatial indicators.

We found that forest and mountain parks, such as Doi Inthanon and Khao Yai, exhibit high semantic diversity and multi-dimensional engagement. In contrast, marine environments like Similan and Koh Chang elicit specialized, high-positivity discourses centered on recreation, while Phu Kradueng presents a unique narrative in which physical effort and adventure shape the emotional responses.

The proposed framework provides a complementary digital approach to traditional survey-based approaches, proving particularly vital during disruptions such as COVID-19. By quantifying the relationship between where tourists go and how they feel, this research offers a scalable tool for the Department of National Parks (DNP) to align ecotourism strategies with national conservation goals, including REDD+ and Blue Carbon initiatives. Ultimately, this insight-driven approach supports the development of more resilient, inclusive, and sustainable management systems for Thailand's natural heritage.

Future research should extend this framework by incorporating independent ecological data, such as satellite-derived habitat heterogeneity indices, vegetation index time series, or species richness databases, to empirically test the relationship between landscape complexity and visitor linguistic richness.

Author Contributions: A.S.: Conceptualization, Methodology, Data analysis, Writing—original draft, Project administration. U.P.: Concept guidance, Review and editing. P.W.: Writing—review support. C.T.N.: Commentary, Review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Srinakharinwirot University, Thailand, as part of a project entitled "The analysis of spatial and movement patterns in travel behavior from social media data." Grant number: SWU326/2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were obtained from publicly available social media platforms, including Flickr and YouTube. Due to platform terms of service and privacy considerations, the raw data cannot be publicly shared. Aggregated data and derived indicators supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the Department of National Parks, Wildlife and Plant Conservation (DNP), Thailand, for providing access to national park boundary data and related spatial information. We also acknowledge the administrative and technical support from the Faculty of Social Sciences, Srinakharinwirot University, JGSEE, King Mongkut's University of Technology Thonburi, and the Faculty of Architecture, Khon Kaen University. During the preparation of this manuscript, the authors used ChatGPT (GPT-5, OpenAI, 2025) to assist in language polishing, improving clarity, and formatting the manuscript. The authors have reviewed and edited all generated content and take full responsibility for the final version of this publication.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
TF-IDF	Term Frequency–Inverse Document Frequency

References

1. Gundersen, V.; Selvaag, S.K.; Junker-Köhler, B.; Zouhar, Y. Visitors' Relations to Recreational Facilities and Attractions in a Large Vulnerable Mountain Region in Norway: Unpacking the Roles of Tourists and Locals. *J. Outdoor Recreat. Tour.* **2024**, *47*, 100807. https://doi.org/10.1016/j.jort.2024.100807.
2. Liu, W.-Y.; Yu, H.-W.; Hsieh, C.-M. Evaluating Forest Visitors' Place Attachment, Recreational Activities, and Travel Intentions under Different Climate Scenarios. *Forests* **2021**, *12*, 171. https://doi.org/10.3390/f12020171.
3. Department of National Parks, Wildlife and Plant Conservation. National Park Act B.E. 2562; Department of National Parks, Wildlife and Plant Conservation: Bangkok, Thailand, 2019. Available online: https://www.reic.or.th/Upload/27_17502_1563854758_53024.pdf (accessed on 10 May 2025).
4. Lin, Y.-X.; Zhan, S.-N.; An, M.-R. Tourism Environment and Activity: Which Matters More in Restoration? *Tour. Recreat. Res.* **2025**, *51*, 586–599. https://doi.org/10.1080/02508281.2025.2495657.
5. Niezgoda, A.; Nowacki, M. Experiencing Nature: Physical Activity, Beauty and Tension in Tatra National Park—Analysis of TripAdvisor Reviews. *Sustainability* **2020**, *12*, 601. https://doi.org/10.3390/su12020601.
6. Schägner, J.P.; Brander, L.; Maes, J.; Paracchini, M.L.; Hartje, V. Mapping Recreational Visits and Values of European National Parks by Combining Statistical Modelling and Unit Value Transfer. *J. Nat. Conserv.* **2016**, *31*, 71–84. https://doi.org/10.1016/j.jnc.2016.03.001.
7. Franěk, M.; Petružálek, J. Perceptual Fluency and Eye Movements When Viewing Urban and Natural Scenes. *Sci. Rep.* **2025**, *15*, 25772. https://doi.org/10.1038/s41598-025-07850-5.
8. Chen, X.; Lee, G.; Joo, D. Natural Landscape Performance: Environmental Restorativeness and Its Influence on Tourist Behavior. *Int. J. Tour. Res.* **2024**, *26*, e2636. https://doi.org/10.1002/jtr.2636.
9. Zheng, X.; Yang, Z.; Fan, Y. Spatial Correlation Mechanism between Natural Landscape Aesthetic Quality and Tourist Perception in Mount Wuyi National Park, China. *Habitat. Int.* **2026**, *168*, 103703. https://doi.org/10.1016/j.habitatint.2025.103703.
10. Wang, J.; Xia, Y.; Wu, Y. Sensing Tourist Distributions and Their Sentiment Variations Using social media: Evidence from 5A Scenic Areas in China. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 492. https://doi.org/10.3390/ijgi11090492.
11. Zhang, G.; Wu, G. Interactive Influence of the Perceived Visual Richness, Greenness and Scenography on Landscape Preference of Urban Woodland. *J. Environ. Psychol.* **2025**, *103*, 102586. https://doi.org/10.1016/j.jenvp.2025.102586.

12. Zheng, T.; Pan, Q.; Huai, S.; Wang, C.; Yan, Y.; Van Eetvelde, V.; Van De Voorde, T. Exploring the Perception of Landscape Elements through User-Generated Data to Support Greenspace Management. *Ecosyst. Health Sustain.* **2024**, *10*, 0282. https://doi.org/10.34133/ehs.0282.
13. Duşcu, D.-M.; Rîşnoveanu, G. Understanding Visitor Preferences: Perceived Importance of Anthropogenic and Natural Forest Features in Supplying Cultural Ecosystem Services. *For. Ecosyst.* **2025**, *13*, 100306. https://doi.org/10.1016/j.fecs.2025.100306.
14. Teles Da Mota, V.; Pickering, C. Geography of Discourse about a European Natural Park: Insights from a Multilingual Analysis of Tweets. *Soc. Nat. Resour.* **2021**, *34*, 1492–1509. https://doi.org/10.1080/08941920.2021.1971809.
15. Flores-Ruiz, D.; Elizondo-Salto, A.; Barroso-González, M.D.L.O. Using Social Media in Tourist Sentiment Analysis: A Case Study of Andalusia during the COVID-19 Pandemic. *Sustainability* **2021**, *13*, 3836. https://doi.org/10.3390/su13073836.
16. Leelawat, N.; Jariyapongpaiboon, S.; Promjun, A.; Boonyarak, S.; Saengtabtim, K.; Laosunthara, A.; Yudha, A.K.; Tang, J. Twitter Data Sentiment Analysis of Tourism in Thailand during the COVID-19 Pandemic Using Machine Learning. *Heliyon* **2022**, *8*, e10894. https://doi.org/10.1016/j.heliyon.2022.e10894.
17. Liu, W.; Wu, L.; Zhao, H. Improving Sentiment Analysis in Tourism through LLM-Enhanced Irony Detection. *Tour. Manag.* **2026**, *112*, 105272. https://doi.org/10.1016/j.tourman.2025.105272.
18. Sitthi, A.; Pimple, U.; Piponiot, C.; Gond, V. Assessing the Effectiveness of Mangrove Rehabilitation Using Above-Ground Biomass and Structural Diversity. *Sci. Rep.* **2025**, *15*, 7839. https://doi.org/10.1038/s41598-025-92514-7.
19. Taecharungroj, V.; Mathayomchan, B. Analysing TripAdvisor Reviews of Tourist Attractions in Phuket, Thailand. *Tour. Manag.* **2019**, *75*, 550–568. https://doi.org/10.1016/j.tourman.2019.06.020.
20. Khruahong, S.; Surinta, O.; Lam, S.C. Sentiment Analysis of Local Tourism in Thailand from YouTube Comments Using BiLSTM. In *Multi-Disciplinary Trends in Artificial Intelligence*; Surinta, O., Kam Fung Yuen, K., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2022; Volume 13651, pp. 169–177, ISBN 978-3-031-20991-8.
21. Ali, T.; Omar, B.; Soulaimane, K. Analyzing Tourism Reviews Using an LDA Topic-Based Sentiment Analysis Approach. *MethodsX* **2022**, *9*, 101894. https://doi.org/10.1016/j.mex.2022.101894.
22. Erdoğan, D.; Kayakuş, M.; Çelik Çaylak, P.; Ekşili, N.; Moiceanu, G.; Kabas, O.; Ichimov, M.A.M. Developing a Deep Learning-Based Sentiment Analysis System of Hotel Customer Reviews for Sustainable Tourism. *Sustainability* **2025**, *17*, 5756. https://doi.org/10.3390/su17135756.
23. Li, K.; Xie, Y.; Liu, Z.; Li, S.; Liu, D. Tourist attraction reviews based on deep learning sentiment analysis system. *JCEIM* **2024**, *14*, 110–113. https://doi.org/10.54097/066t1j26.
24. Shaik, T.; Tao, X.; Dann, C.; Xie, H.; Li, Y.; Galligan, L. Sentiment Analysis and Opinion Mining on Educational Data: A Survey. *Nat. Lang. Process. J.* **2023**, *2*, 100003. https://doi.org/10.1016/j.nlp.2022.100003.
25. Hausmann, A.; Toivonen, T.; Fink, C.; Heikinheimo, V.; Kulkarni, R.; Tenkanen, H.; Di Minin, E. Understanding Sentiment of National Park Visitors from Social Media Data. *People Nat.* **2020**, *2*, 750–760. https://doi.org/10.1002/pan3.10130.
26. Visariya D, Prajapati P, Bhatt, N. YouTube comment sentiment analysis using NLP approach. *AIP Conf Proc.* **2025**, *3255*, 020005. https://doi.org/10.1063/5.0254716.
27. Gholamnia, M.; Eslamirad, N.; Sajadi, P.; Zarin, Z.; Pilla, F. Leveraging Large Language Models for Citizen-Centric Urban Accessibility Analysis: A Case Study Using Airbnb Reviews in Dublin. *Spat. Inf. Res.* **2025**, *33*, 43. https://doi.org/10.1007/s41324-025-00646-9.
28. Khamphakdee, N.; Seresangtakul, P. An Efficient Deep Learning for Thai Sentiment Analysis. *Data* **2023**, *8*, 90. https://doi.org/10.3390/data8050090.
29. Kiran Kumar, B.S.; Prajwal, M.L.; Nivedita. Sentiment Analysis of Indian Tourist Place Reviews: A Machine Learning-Based Exploration. In *Proceedings of the 2023 4th International Conference on Intelligent Technologies (CONIT)*; IEEE: Bangalore, India, 2024; pp. 1–5.
30. Kovacs-Györi, A.; Ristea, A.; Kolcsar, R.; Resch, B.; Crivellari, A.; Blaschke, T. Beyond Spatial Proximity—Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 378. https://doi.org/10.3390/ijgi7090378.
31. Fox, N.; August, T.; Mancini, F.; Parks, K.E.; Eigenbrod, F.; Bullock, J.M.; Sutter, L.; Graham, L.J. "Photosearcher" Package in R: An Accessible and Reproducible Method for Harvesting Large Datasets from Flickr. *SoftwareX* **2020**, *12*, 100624. https://doi.org/10.1016/j.softx.2020.100624.
32. SmugMug, Inc. *Flickr APIs Terms of Use*; Available online: https://www.flickr.com/help/terms/api (accessed on 1 April 2025).
33. Google LLC. YouTube API Services—Terms of Service; Google LLC. Available online: https://developers.google.com/youtube/terms/api-services-terms-of-service (accessed on 1 April 2025).

34. Shen, H.; Aziz, N.F.; Liu, J.; Huang, M.; Yu, L.; Yang, R. From Text to Insights: Leveraging NLP to Assess How Landscape Features Shape Tourist Perceptions and Emotions toward Traditional Villages. *Environ. Res. Commun.* **2024**, *6*, 115006. https://doi.org/10.1088/2515-7620/ad8ca3.
35. Benoit, K.; Watanabe, K.; Wang, H.; Nulty, P.; Obeng, A.; Müller, S.; Matsuo, A. Quanteda: An R Package for the Quantitative Analysis of Textual Data. *JOSS* **2018**, *3*, 774. https://doi.org/10.21105/joss.00774.
36. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *JOSS* **2019**, *4*, 1686. https://doi.org/10.21105/joss.01686.
37. Wickham, H. Stringr: Simple, Consistent Wrappers for Common String Operations. 2023. Available online: https://rdrr.io/cran/stringr/ (accessed on 1 April 2025).
38. Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D. A Grammar of Data Manipulation, R Package Version 1.1.4; Posit Software, PBC: Boston, MA, 2025. Available online: https://CRAN.R-project.org/package=dplyr (accessed on 1 April 2025).
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-Learn: Machine Learning in Python. *arXiv* **2012**, arXiv:1201.0490. https://doi.org/10.48550/arXiv.1201.0490.
40. Phatthiyaphaibun, W.; Chaovavanich, K.; Polpanumas, C.; Suriyawongkul, A.; Lowphansirikul, L.; Chormai, P.; Limkonchotiwat, P.; Suntorntip, T.; Udomcharoenchaikit, C. PyThaiNLP: Thai Natural Language Processing in Python. In Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), Singapore, 6 December 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 25–36.
41. Rehurek, R.; Sojka, P. *Gensim–Python Framework for Vector Space Modelling*; NLP Centre, Faculty of Informatics, Masaryk University: Brno, Czech Republic, 2011; Volume 3.
42. Seabold, S.; Perktold, J. Statsmodels: Econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference; 2010 Jun 28–Jul 3; Austin, TX, USA. Austin (TX): SciPy; 2010. p. 92–96. https://doi.org/10.25080/Majora-92bf1922-011.
43. Bird, S.; Loper, E. NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions; 2004 Jul; Barcelona, Spain. Association for Computational Linguistics; 2004. p. 214–217. Available from: https://aclanthology.org/P04-3031/.
44. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
45. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. https://doi.org/10.1162/COLI_a_00049.
46. Lowphansirikul, L.; Polpanumas, C.; Jantrakulchai, N.; Nutanong, S. WangchanBERTa: Pretraining Transformer-Based Thai Language Models. *arXiv* **2021**, arXiv:2101.09635.
47. Garrido-Merchan, E.C.; Gozalo-Brizuela, R.; Gonzalez-Carvajal, S. Comparing BERT Against Traditional Machine Learning Models in Text Classification. *JCCE* **2023**, *2*, 352–356. https://doi.org/10.47852/bonviewJCCE3202838.
48. Spellerberg, I.F.; Fedor, P.J. A Tribute to Claude Shannon (1916–2001) and a Plea for More Rigorous Use of Species Richness, Species Diversity and the 'Shannon–Wiener' Index. *Glob. Ecol. Biogeogr.* **2003**, *12*, 177–179. https://doi.org/10.1046/j.1466-822X.2003.00015.x.
49. Oksanen, J.; Simpson, G.L.; Blanchet, F.G.; Kindt, R.; Legendre, P.; Minchin, P.R.; O'Hara, R.B.; Solymos, P.; Stevens, M.H.H.; Szoecs, E.; et al. Vegan: Community Ecology Package 2001, Version 2.7-2; Comprehensive R Archive Network (CRAN): Vienna, Austria, 2001. Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR; et al. vegan: Community Ecology Package. R package version 2.7-2. 2025. Available from: https://CRAN.R-project.org/package=vegan.
50. Luo, J.M.; Shang, Z. Hiking Experience Attributes and Seasonality: An Analysis of Topic Modelling. *Curr. Issues Tour.* **2024**, *27*, 2984–3000. https://doi.org/10.1080/13683500.2023.2245955.
51. Al Jassim, R.S.; Al Mansoory, S.; Al-Dhuhouri, G.Z.S. Strategic Insights from Customer Feedback: Study of Hotel Reviews Using Logistic Regression. In *Proceedings of the 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*; IEEE: Vallette, Malta, 2024; pp. 785–790.
52. Sharaf, L. Sentiment Analysis of Twitter Data Using TF-IDF and Logistic Regression. SSRN 2025. https://doi.org/10.2139/ssrn.5505839 .
53. Zhao, J.; Abdul Aziz, F.; Song, M.; Zhang, H.; Ujang, N.; Xiao, Y.; Cheng, Z. Evaluating Visitor Usage and Safety Perception Experiences in National Forest Parks. *Land* **2024**, *13*, 1341. https://doi.org/10.3390/land13091341.
54. Alexandridis, K. Assessing Cognitive and Social Attitudes toward Environmental Conservation in Coral Reef Social-Ecological Systems. *Soc. Sci.* **2018**, *7*, 109. https://doi.org/10.3390/socsci7070109.
55. Hasson, F.; Keeney, S.; McKenna, H. Research Guidelines for the Delphi Survey Technique. *J. Adv. Nurs.* **2000**, *32*, 1008–1015.

56. Fleiss, J.L. Measuring Nominal Scale Agreement among Many Raters. *Psychol. Bull.* **1971**, *76*, 378–382. https://doi.org/10.1037/h0031619.
57. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159. https://doi.org/10.2307/2529310.
58. Ye, C.; Chen, Z.; Ding, Z. How Does High Temperature Weather Affect Tourists' Nature Landscape Perception and Emotions? A Machine Learning Analysis of Wuyishan City, China. *PLoS ONE* **2025**, *20*, e0323566. https://doi.org/10.1371/journal.pone.0323566.
59. Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.
60. Üzülmez, M.; Ercan İştin, A.; Barakazı, E. Environmental Awareness, Ecotourism Awareness and Ecotourism Perception of Tourist Guides. *Sustainability* **2023**, *15*, 12616. https://doi.org/10.3390/su151612616.
61. Xiao, W. Theme Modeling Study of Online Media Data for Ecotourism Based on Natural Language Processing. In *Proceedings of the 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE)*; IEEE: Shenyang, China, 2025; pp. 1009–1016.
62. Ma, B.; Zeng, W.; Xie, Y.; Wang, Z.; Hu, G.; Li, Q.; Cao, R.; Zhuo, Y.; Zhang, T. Boundary Delineation and Grading Functional Zoning of Sanjiangyuan National Park Based on Biodiversity Importance Evaluations. *Sci. Total Environ.* **2022**, *825*, 154068. https://doi.org/10.1016/j.scitotenv.2022.154068.
63. Abarca, H.; Morán-Ordoñez, A.; Villero, D.; Guinart, D.; Brotons, L.; Hermoso, V. Spatial Prioritisation of Management Zones in Protected Areas for the Integration of Multiple Objectives. *Biodivers. Conserv.* **2022**, *31*, 1197–1215. https://doi.org/10.1007/s10531-022-02383-z.
64. Cvetković, M.; Brankov, J.; Ćurčić, N.; Pavlović, S.; Dobričić, M.; Tretiakova, T.N. Protected Natural Areas and Ecotourism—Priority Strategies for Future Development in Selected Serbian Case Studies. *Sustainability* **2023**, *15*, 15621. https://doi.org/10.3390/su152115621.
65. Chang, R.; Yang, D.; Lyu, D.; Cao, J.; Wei, Y.; Xiao, D. The Linguistic Feedback of Tourism Robots Significantly Influences Visitors' Ecotourism Behaviors. *Sci. Rep.* **2025**, *15*, 16015. https://doi.org/10.1038/s41598-025-01045-8.