Applied Decision Analytics

Journal homepage: www.ada-journal.org
ISSN: 3104-2945

Machine Learning–Driven Sentiment Analysis for Tourism Insights: The Case of North-East Indian States

Sandeep Bhattacharjee[1,*]

[1] Amity School of Business, Amity University Kolkata, Kolkata, West Bengal, India

ARTICLE INFO

Article history:
Received 29 July 2025
Received in revised form 21 October 2025
Accepted 10 November 2025
Available online 26 December 2025

Keywords:
Emotion Detection, Feature Extraction, Opinion Mining, Polarity Classification, Sentiment

ABSTRACT

This study is related to online reviews- publicly available from the northeastern states of India, that includes Sikkim, Meghalaya, Arunachal Pradesh, Assam, Nagaland, Mizoram, and Manipur. In this paper, we evaluate the impact of sentiment analysis for improving tourism decision-making. It addresses the limitations of traditional surveys for capturing emotional and informal experiential insights. Open-access datasets from TripAdvisor, Reddit, and travel blogs (GitHub) were systematically tokenized, lemmatized, and stop-word-free text was used to derive clean and interpretable textual input. NRC Emotion Lexicon was used for understanding the emotional tone, while VADER and TextBlob derived sentiment polarity, resulting in the generation of a state-wise sentiment index. The K-means clustering method helped in the segmentation of Emotion and experience themes. Results reveal a favourable attitude towards Sikkim and Meghalaya due to natural beauty and hospitality, whereas Assam and Nagaland reflected mixed sentiments due to infrastructure limitations. Dominant feelings of Joy, trust, and surprise were associated with positive sentiments, while sadness and dislike were related to service issues. The study determined the role of AI-driven sentiment analytics for enhancing destination branding, service improvement, and policy planning, which can support a strong foundation for data-based decision making for the growth of tourism in these regions.

1. Introduction

Sentiment analysis, in plain words, denoted as "the computational study of opinions, emotions, and attitudes stated in text" [1], has evolved as a critical tool across various domains. In other words, [2] regarded it as the detection of sentiment polarity and subjectivity, while others [3] related it to real-world marketing applications and exactness. Another researcher [4] discussed its evolution by stating the primary problems associated with its application. Feature extraction [5] and syntax-content models [6] advanced enhanced classification, and performance were further associated with embeddings and emotion integration [7, 8]. Some of the comparative studies [9] and unsupervised detection [10] exhibited methodological advances in this direction. Some of these applications include social media[11], healthcare [12], finance [13], and tourism [14]. Empirical research [15-17]

* *Corresponding author.*
E-mail address: sandeepbitmba@gmail.com

https://doi.org/10.66972/ada21202613

© The Author(s) 2026 | Creative Commons Attribution 4.0 International License

also highlights the relevance of sentiment analysis for the interpretation of perceptions in from the perspective of tourism in North-East India.

Sentiment analysis can be used for recording, understanding, and evaluating human views, opinions, and experiences for social, commercial, and other contexts [18, 19, 1]. Other studies [20] demonstrate its relevance in marketing, whereas others [21, 22] applied it for evaluating neighborhood and environmental perceptions, respectively. In addition, experts also asserted its relevance to tourism [23, 14]. In terms of methodologies, advancements include feature extraction [24], ontology-based methodologies [25-27], multimodal strategies[28], embeddings [29], classifier ensembles [30], and affective computing [19, 31]. From a business perspective, Sentiment analysis is becoming relevant across various domains, such as advertising, social media, e-learning, and energy and climate [32-35]. Thus, Sentiment analysis acts as a tool of interpretation of tourist attitudes, experiences, and fulfills the need for data-driven decision-making in the travel industry. In addition, experts [36] also reviewed conceptual and methodological developments in hospitality and tourist sentiment analysis, while others [14] highlighted the importance of big data for destination management. In addition, others [37] applied semantic clustering for tourism recommendation systems and evaluated automated sentiment analysis approaches, discovering methodological variations in its usage[38]. It was observed that feature-based sentiment analytics were related to negative reputations and tourist experiences for different regions [39, 40]. Social media-based sentiment research also recognized tourists' attitudes during COVID-19 [41, 42] and air quality concerns [43]. More detailed research revealed applications such as TripAdvisor and Twitter analytics that measured loyalty, engagement, and decision-making [44-47], besides fuzzy framework models that labeled travel safety insights [48]. While some experts [49] pointed to user-generated content as the main source of knowledge for decision making, others [50] pointed to its platform-specific implications. These studies emphasize the significance of sentiment analysis for tourism planning, marketing, and tourist experience management.

1.1 Motivation for selecting North-East India as a case study

The northeastern states of India are known for their richness in cultural diversity, natural landscapes, and undiscovered destinations, and remain unexplored as favored tourism destinations. No prior studies have been conducted in these regions, although it provides an excellent avenue for tourism development. The present work proposes using user-generated content from internet platforms and social media to support destination management strategies, marketing strategies, and policy development exclusively for these attractive regions. Furthermore, sentiment analysis can serve as an efficient tool for developing solutions in terms of accessibility, sustainability, and the promotion of region-specific attractions of the northeastern states of India.

1.2 Contribution of the current research

The present work serves as a step forward in meeting the goals of promoting northeastern Indian tourism by using sentiment analysis for the development of region-specific insights. In this respect, section 2 addresses contextual and cultural variations using the integrated tools of text mining, NLP, opinion mining, polarity detection, emotion classification, and aspect-based techniques. In Section 3, machine learning and deep learning methods comprising advanced neural architectures (LSTM, BERT, FinBERT) and conventional models (Naïve Bayes, SVM, Random Forest) were evaluated utilizing reliable performance metrics. Section 4 implements the methodology to the North-East India case study, gathering data from tourist reviews, social media, and travel platforms, initial processing,

feature engineering, and model application. Section 5 discusses analysis research methodologies stagewise. Section 6 involves sentiment score, polarity, emotional patterns, and visualization analysis results, while Section 7 discusses the main findings and related discussions. Section 8 addresses sentiment analysis's effects on tourism policy and practices, and Section 9 concludes with recommendations.

2. Sentiment Analysis: Concepts and Framework

2.1 Definition and theoretical foundations of sentiment analysis

Sentiment analysis, also known as opinion mining, is a part of natural language processing that identifies, extracts, and classifies subjective information from textual data [15]. Research [19] often associates sentimental analysis with the assessment of emotional intensity in different languages and exhibits sentiment polarity (positive, negative, or neutral). Sentiment analysis is domain specific analyses of user-generated content, such as reviews, social media posts, and comments, which capture public opinion and emotional trends.

It uses linguistic principles besides syntactic and semantic analysis for meaning extraction and emotions expressed through language [51]. Approaches such as part-of-speech tagging and dependency parsing help in the identification of emotional lexicon. Sentiment analysis has its fundamental roots in emotional science, which explores human emotions and attitudes [52]. It begins with the text as an input reflecting the writer's emotional or evaluative state and then integrates computational models with emotion-based psychological theories for deriving patterns for generalization. This analysis uses both machine learning and statistical learning techniques to reveal sentiment patterns from a computational perspective [53]. The prior methods were lexicon-based, which are now transformed into deep learning and transformer-based models [54]. In this context, clues related to Social, cultural, and contextual norms, such as idioms, sarcasm, and cultural metaphors, jointly influence sentimental perception. Therefore, the meanings of cultural and contextual aspects are needed for emotion modeling. In this regard, emotional indicators found in textual communication exhibit group sentiments or perceptions, necessary for predictive analytics in marketing, politics, and tourism.

2.2 Text mining, natural language processing, and opinion mining techniques

Text mining refers to the integrated approach that involves statistical and computational techniques to identify significant patterns and insights in unstructured textual material. It reveals hidden patterns in large text databases by the use of data mining, information retrieval, and linguistic processing [55, 56]. Some of the common utilities include social media analytics, patent discovery, and the study of scientific literature [57]. Evidence can be extracted using text mining, and new trends in emotions can also be discovered with automated information extraction [58]. Text mining has evolved from natural language processing(NLP) and understands, interprets, and generates human language for spoken language and written data [59, 60]. BERT and GPT are some of the latest advancements that have significantly improved language development and contextual understanding [61, 62]. Some of the most popular NLP applications include healthcare, industrial automation, and information systems, which convert raw textual data into meaningful insights [63, 64].

2.3 Polarity detection, emotion classification, and aspect-based sentiment analysis

Aspect-based Sentiment Analysis (ASBA) has evolved as a major approach for sentiment analysis, enabling the understanding of attitudes toward specific qualities in each text and the identification of complex sentiment expressions through lexical, syntactic, and semantic approaches [65]. Deep learning architectures like BERT, GRU, and Graph Convolutional Networks enable aspect extraction and polarity classification in multilingual environments. It also deals with comprehension and aspect detection with advancements in aspect extraction and polarity classification in multilingual environments. Adaptive and domain-independent lexicon frameworks were further developed for contextual comprehension and aspect detection [66, 10]. Topic modeling and semantic augmentation are hybrid models that improve interpretability [67]. These models were used for reputation management and behavioral insights in the travel and hospitality industries [68, 69], although concerns regarding domain adaptation and aspect granularity continue to exist [70].

Context-based review evaluation is related to sentiment polarity in different areas, whereas emotions are related to linguistic, cultural, and situational factors. Both analyses are derived from user content in social media, dialogues, and marketing campaigns as seen during COVID-19 discussions, Setswana texts, diverse digital marketing contexts, and other cross-cultural methodologies [71, 72]. Recent reviews reflect the use of deep learning methods, generalization issues, and modern approaches for scalable, context-sensitive sentiment models [73, 74]. Emotion-based theoretical studies further support the understanding of emotional reactions and model development [75]. These literatures highlight the potential of context-aware sentiment analysis in integration with artificial intelligence for accurate, interpretable, and culturally responsive results.

2.4 Machine Learning Techniques for Sentiment Analysis

Both machine learning(ML) and deep learning utilize advanced techniques for pattern recognition, forecasting, and decision-making across diverse fields. While standard traditional machine learning algorithms can be used for structured data analysis, deep learning models are used for automatic extraction of hierarchical features for large-scale text classification, image analysis, and drug discovery [76-78], using more computing power [79, 80]. Some applications of deep learning (DL) include medical diagnostics, intrusion detection, and material property prediction [81, 82]. ML and DL can also be integrated for model selection, optimization, and interpretability, which suggests a shift in momentum toward intelligent, data-driven solutions [83].

Random Forests, Support Vector Machines (SVM), and Naïve Bayes are some of the traditional tools for regression, classification, and predictive analytics. Naïve Bayes classifiers deal with text classification, intrusion detection, and sentiment analysis, with additive features of selective, non-parametric, and hidden multiclass variants for improvement in performance in text classification, image analysis, and drug discovery [84, 85,86]. Other advanced features with SVM solvers ,smooth SVMs, Kernel methods are also available for pattern recognition, image classification, feature selection and margin maximization[87]. Feature learning, optimization strategies, and random forests acts as ensemble learners for applications of groundwater mapping, bioinformatics, and remote sensing [88-90].

These models act as basic standards for modern machine learning and deep learning applications that include advanced neural architectures, such as Long Short-Term Memory (LSTM) networks, BERT, and FinBERT.LSTM networks were developed for capturing long term dependencies and managing gradient problems in recurrent neural networks, as observed during time-series forecasting, energy consumption prediction, hydrology, and emergency event detection [91-94].

Another pre-trained model known as BERT has been used for text categorization, finding fake news, recognizing sarcasm, and aspect-based sentiment analysis in the sectors of biomedical research, social media, and cybersecurity [95, 96]. In Finance, it serves as a tool for predicting market sentiments, analyzing financial texts, and assessing risk [97,98]. These models have a significant impact on both general-purpose and domain-specific natural language processing tasks.

2.5 Performance metrics and model evaluation criteria

In the present work, the Sentiment Polarity Scores (Positive, Neutral, Negative) indicator was used to measure overall tourist satisfaction and emotional tone of reviews during sentiment analysis.

Similarly, the percentage distribution of sentiment classes by state indicator compares regional sentiment trends and identifies state-level differences during aggregated sentiment distribution analysis. The Topic Coherence Score indicator reveals the significance of the concepts and the extracted keywords using the topic modeling algorithm. The most dominant topics per state serve as a metric for identifying primary visitor concerns, interests, and attractions. It underlines tourist experiences unique to each state, including eco-tourism, cultural heritage, festivals, and adventurous activities. The Silhouette Score measures distance within and between clusters using K-means clustering analysis. These variables support the analysis of the variance by evaluating web-scraped reviews from seven Northeastern states and identifying patterns of satisfaction, destination image, and emotional tone. Thematic alignment of clusters integrates data-driven clusters with relevant tourism categories during cluster analysis. These indicators enable the identification of gaps by evaluating web-scraped reviews from seven Northeastern states and determining patterns of satisfaction, destination image, and emotional tone.

This work helps to fulfill the objectives of (1) measuring sentiment polarity across states, (2) identifying dominant emotions and visitor-related problems, and (3) segmenting locations based on review patterns. It also aligns results with proposed actionable insights for tourism policy, marketing, and sustainable destination development.

3. Case Study: Sentiment Analysis of Tourism in North-East India

In this study, we select the Northeastern states of India as they offer distinct ecological, cultural, and socioeconomic qualities distinct from other tourism destinations within the country.

The region includes Sikkim, Assam, Meghalaya, Arunachal Pradesh, Manipur, Mizoram, and Nagaland, is well-known for its biodiversity, beautiful landscapes, and rich cultural legacy, offering a diverse range of activities from eco-adventure and rural tourism to cultural festivals and tribal heritage.

North-eastern states have not yet been explored in terms of tourism and remain underdeveloped. Challenges such as infrastructural limitations, seasonal accessibility problems, and limited digital promotion dominate the underdevelopment of these tourism destinations. State-specific variations may exist and need to be explored to generate valuable insights for ensuring the growth of these regions in different proportions.

The present work is motivated to uncover the knowledge in this domain using relevant methods, which include sentiment analysis, topic modeling, and clustering.

Preliminary research indicates multiple data sources of views and opinions of tourists returning from these destinations. These sources include tourist reviews, social media platforms, travel websites, and other open-access datasets. Tourist reviews include feedback from travelers available on tourist review websites such as TripAdvisor, which provides detailed information on lodging,

activities, and services, with different degrees of satisfaction. New perspectives on thoughts, experiences, destination preferences, and emotional reactions can also be gathered from social media posts and comments on websites such as Reddit, travel blogs, etc. The data collected from these sources [99] provides the initial dataset for further analysis and discussion. This dataset consists of both structured and unstructured content, which can be analyzed for sentiment generation, thematic interests grouping, and state-specific patterns. Further, we propose to use feature engineering to ensure the high quality of input data for better analytical results. We also need to transform the text content to lowercase to preserve consistency. Semantic consistency is achieved using tokenization and lemmatization. Sentiment analysis on the processed dataset was represented using numerical sentiment features as subjectivity metrics and polarity scores. Dominant terms in the text with higher TF-IDF can be extracted from the document-term matrix. Visualization charts such as Radar charts, Parallel coordinates plots, and t-SNE Projection reveal insights on state-based dominance of facts and are suitable for inter-state comparison. The current research discusses these tourism patterns and discovers a rich and robust framework in reflecting the complexities of tourist experiences.

4. Methodology

4.1 Method used

The investigation on the tourism dataset was conducted stagewise:

Stage 1: In the first stage, we extracted tourism-related text data from TripAdvisor, Reddit, travel blogs, and other social media sources and stored them in an open-access dataset [99].

Stage 2: In the next stage, we cleaned and normalized the raw text and converted all the text content to lowercase. This was immediately followed by splitting text into words or tokens, where

$$Tokens = split(Review) \tag{1}$$

Further, stop-word removal was initiated, with

$$T' = T - S \tag{2}$$

where T= total tokens, S= stopword set.

This was preceded by the lemmatization of transformed words to their root forms, wherein terms such as “travelling” and “traveled” were regarded as “travel”.

Stage 3: In this stage, two principal feature representations were applied:

i. TF-IDF (Term Frequency–Inverse Document Frequency):

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right) \tag{3}$$

where $TF(t, d)$= term frequency in document d,
$DF(t)$= number of documents containing the term t ,
N= total number of documents.

ii. Word Embeddings (Word2Vec/BERT):

Each token was mapped to a dense vector $\mathbf{v_i} \in \mathbb{R}^n$, enabling semantic representation

Stage 4: Sentiment scoring was represented by Sentiment polarity (P) quantifies the positivity or negativity of a review.

i. Lexicon-based methods (e.g., VADER/TextBlob):

Each word w_i has an assigned polarity score $s(w_i) \in [-1, +1]$.

The overall sentiment polarity for a review $R$is:

$$PR = \sum_{i=1}^{n} s(w_i) \quad (4)$$

where $P_R > 0$→ Positive, $P_R < 0$→ Negative, $P_R = 0$→ Neutral.

ii. Compound Sentiment Score (VADER):

$$CR = \frac{\sum s(w_i)^2 + \alpha}{\sum s(w_i)} \quad (5)$$

with $\alpha = 15$(normalization factor).

Stage 5: Beyond polarity, emotion classes (e.g., joy, anger, surprise) were extracted using NRC Emotion Lexicon.

For review r, We computed :

$$\text{Etype}(R) = \frac{\text{Total Words}(R)}{\text{Count}_{\text{emotion}}(R)} \quad (6)$$

allowing computation of emotional intensity per review.

Stage 6: For Aggregation and State-wise Analysis, State-level sentiment index (SI_{state}) was computed as:

$$SI_{\text{state}} = \sum_{j=1}^{m} PR_j \quad (7)$$

where m= total reviews for a given state.

The aggregation resulted in a comparative sentiment covering Sikkim, Assam, Meghalaya, Arunachal Pradesh, Nagaland, Mizoram, and Manipur.

Stage 7: K-Means was used for thematic and emotional groupings for clustering and visualization.

$$minJ = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (8)$$

where C_i= cluster, μ_i= cluster centroid. Results were visualized using word clouds, polarity distribution plots, and sentiment heatmaps.

Stage 8: For model evaluation,

Performance metrics (Accuracy, Precision, Recall, and F1-score) were computed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

Therefore, this multi-stage sentiment analysis pipeline enabled a structured, explainable, and comparative understanding of tourist satisfaction, emotional tone, and destination image across North-Eastern India (Figure 1).

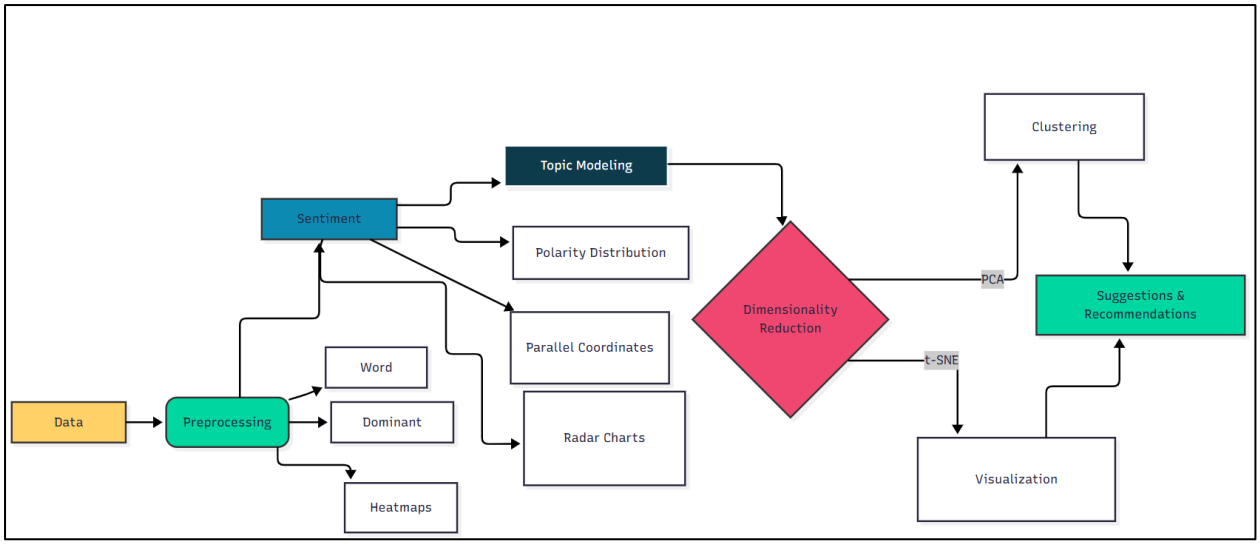


Fig. 1. Tourist review sentiment pipeline(*Source: Author Analysis*)

5. Results and Discussion

5.1 State-Wise Word Cloud

The word cloud visualizations for tourist evaluations across the seven North Eastern states of India (Sikkim, Mizoram, Manipur, Nagaland, Arunachal Pradesh, Assam, and Meghalaya) reveal common lexical patterns based on terms such as "tour," "group," "5," "adult," and "TripAdvisor." Tourists appreciate group travel and ratings-focused chat, as per language classification.. State-specific terms like “Gangtok” in Sikkim and “Shillong” in Meghalaya indicate local visitor interests. These findings prepared the stage for attitude analysis and regional tourism planning (Figure 2).

Fig. 2. Word cloud for seven states (Source: Author analysis)

5.2 State-Wise Dominant Words and Heatmap

Tourist evaluations in Mizoram, Sikkim, Arunachal Pradesh, Meghalaya, Nagaland, and Manipur commonly included travel-related keywords like “trip,” “group,” “tour,” and “experience.” Region-centric attractions are emphasized by place-specific references, such as "Gangtok" in Sikkim and

"Aizawl" in Mizoram. This identical linguistic pattern across states points to a similar tourism narrative, providing the framework for comparative sentiment analysis and destination-specific marketing strategies for Northeast India (Table 1).

Table 1
Dominant words for Indian Northeastern states

Words	Arunachal	Assam	Manipur	Meghalaya	Mizoram	Nagaland	Sikkim
Top 1	tour	tour	tour	tour	Tour	tour	sikkim
Top 2	group	group	group	group	Visit	group	tour
Top 3	visit	adult	visit	tripadvisor	Tripadvisor	visit	group
Top 4	nights	hyderabad	adult	shillong	honeymoon	adult	visit
Top 5	honeymoon	price	price	visit	Family	price	gangtok
Top 6	patna	varies	varies	adult	Packages	varies	tripadvisor
Top 7	family	visit	toursfromâ, [1]per	price	Reach	shillong	lake
Top 8	pradesh	toursfrom, çπper	tripadvisor	varies	Mizoram	meghalaya	places
Top 9	india	tours	meghalaya	bubbles	Group	toursfromâ,[1]per	adult
Top 10	packages	city	shillong	meghalaya	India	tripadvisor	darjeeling
Top 11	arunachal	tripsfrom, çπper	packages	tours	Travel	honeymoon	india
Top 12	trip	temple	family	honeymoon	contributions	packages	pelling

Source: Calculated by author

Terms such as “trip,” “tour,” “travel,” and “experience” were commonly identified for all the Indian Northeastern states. State-level lexical terms such as “Gangtok” for Sikkim and “Aizawl” reflect the most favored destinations of visit based on tourist review text contents. Major high-frequency clusters in Assam and Sikkim suggest an increase in tourism activity or content volume in text contents. Results exhibit linguistic evidence for both shared and distinctive experiential descriptions that can be used for planning destination marketing and regional tourism initiatives (Figure 3).

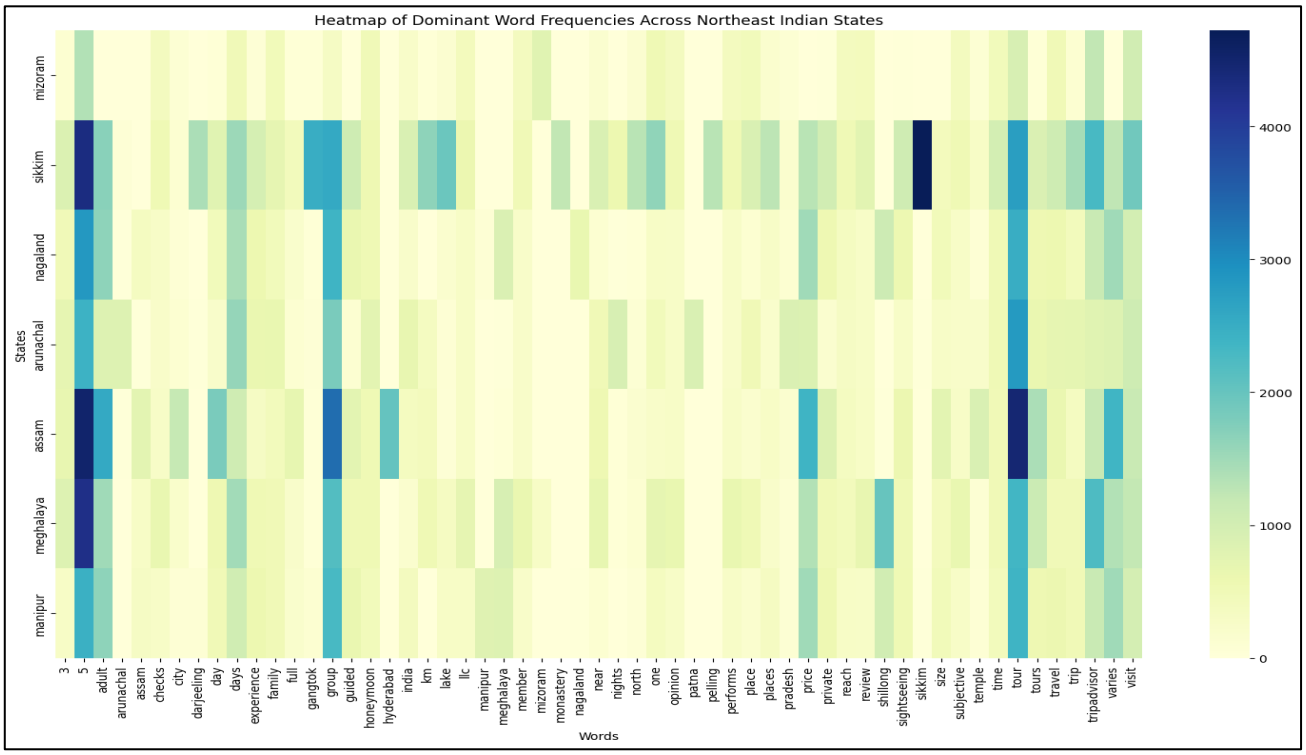
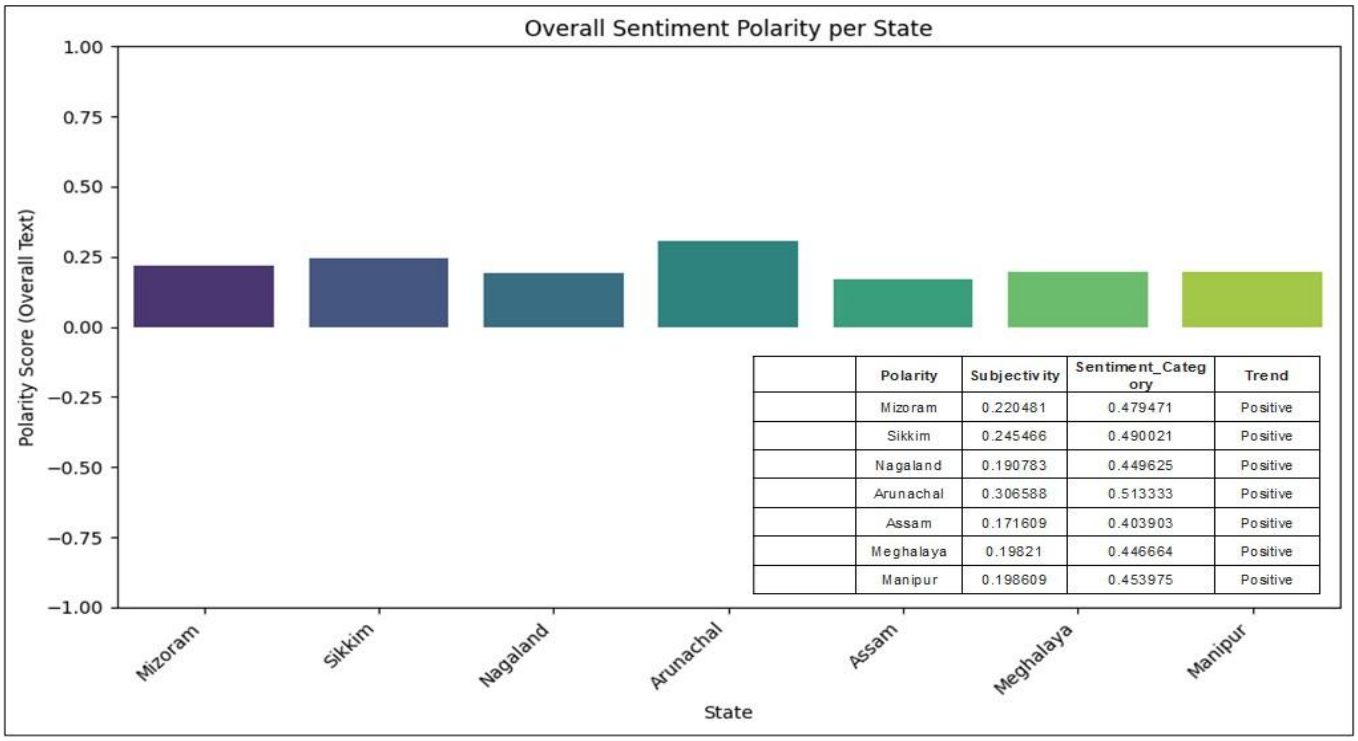


Fig. 3. Heatmap of Dominant Words (Source: Author analysis)

5.3 Polarity Score Distribution (Histogram or Violin Plot)

Polarity scores reflect the range of reviews in positive and negative sides. Consistent favorable trends were observed (Assam-Arunachal Pradesh: 0.17-0.31). Most of the sentiments observed were positive, highlighting the positive experiences among tourists. This trend clearly indicates the optimism among tourists for these destinations. These trends can be utilized for planning and decision-making in destination branding strategies for this region in India (Figure 4).



	Polarity	Subjectivity	Sentiment_Category	Trend
	Mizoram	0.220481	0.479471	Positive
	Sikkim	0.245466	0.490021	Positive
	Nagaland	0.190783	0.449625	Positive
	Arunachal	0.306588	0.513333	Positive
	Assam	0.171609	0.403903	Positive
	Meghalaya	0.19821	0.446664	Positive
	Manipur	0.198609	0.453975	Positive

Fig. 4. Overall sentiment polarity per state (Source: Author analysis)

5.4 Parallel Coordinates Plot

The review length, polarity, and subjectivity are the three observed normalized sentiment variables for tourist reviews across seven states in Northeastern India (Figure 4). Arunachal Pradesh exhibits intense emotions and favorable sentiments with major variations between personal and professional interpretation. In contrast, reviews for Assam and Nagaland showed a trend towards objective or impartial narratives with a lower level of favorability. High positive perception among toursits for these states suggests the scope of sustainable tourism through effective destination branding (Figure 5).

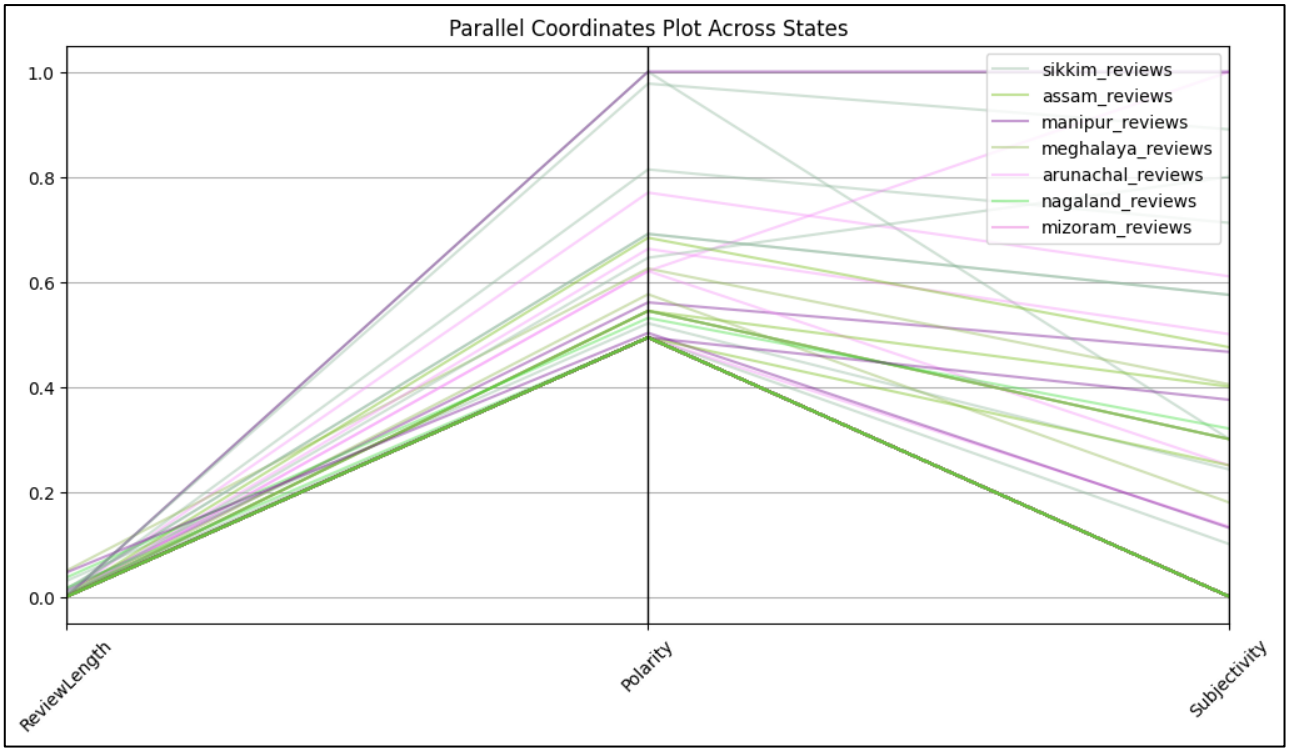


Fig. 5. Parallel coordinates plots (Source: Author Analysis)

5.5 State-wise Emotion using Radar chart

Common use of phrases such as “tour,” “group,” and “bubbles” reflects a growing trend among reviewers for organized travel and memorable experiences. These words highlight the impact of planned tours and travel on customer satisfaction and customer delight(Figure 6).

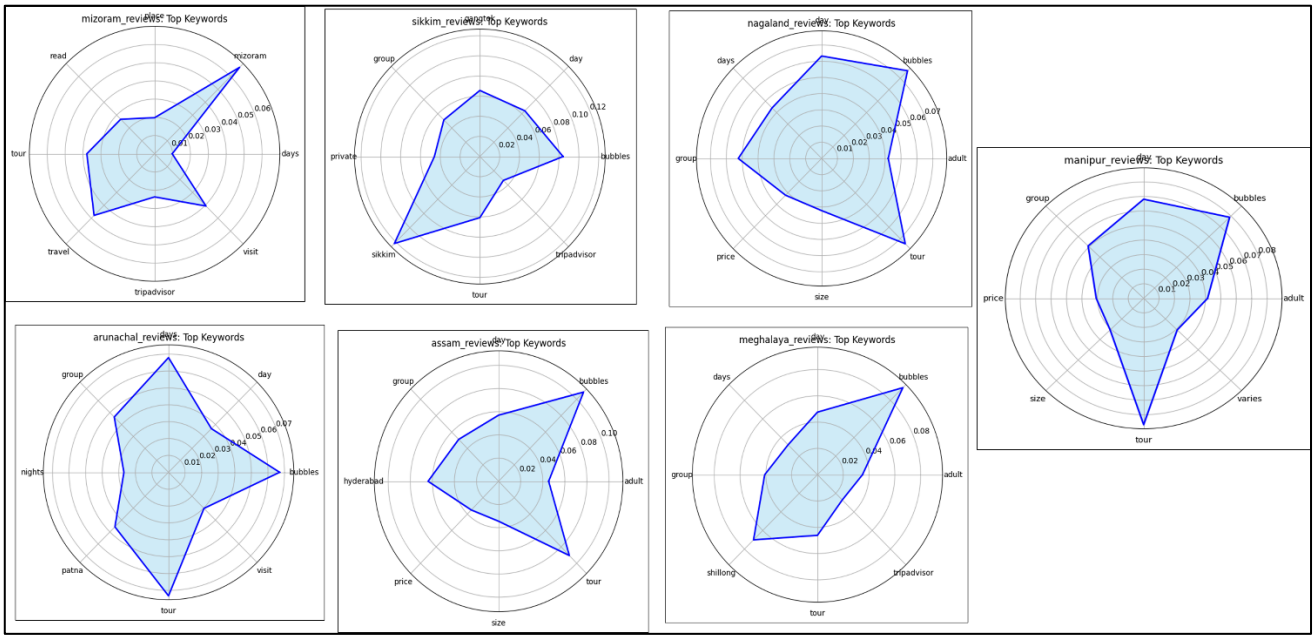


Fig. 6. Radar Charts showing state-wise Emotions (Source: Author Analysis)

5.6 Hierarchical Clustering Dendrogram

Linguistic and experience-based segmentation reflect distinct categories of experiences as suggested by the reviews of visitors for these states. While Arunachal Pradesh and Mizoram are grouped under cultural-linguistic preferences, Assam demonstrates the greatest lexical variation. Major reasons for variation can be diverse tourist demographics or unique factors like pricing and adult-oriented experiences. Sikkim and Nagaland, grouped, reflect differences from the central cluster. This analysis reveals the interstate differences in review semantics and tourist conversations (Figure 7).

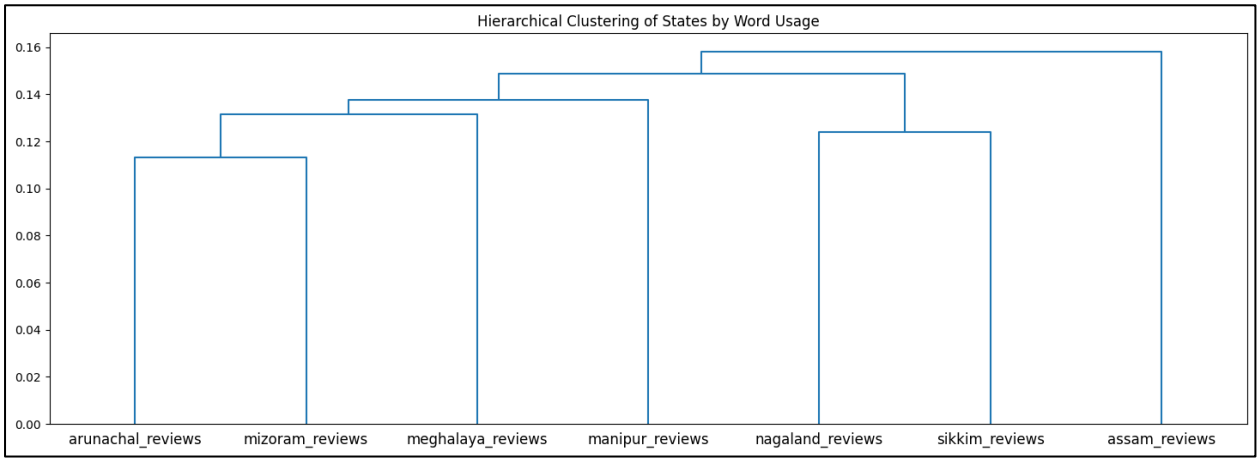


Fig. 7. Hierarchical clustering of Indian northeastern states (Source: Author analysis)

5.7 t-*SNE Projection of Reviews*

The t-SNE projection represents a two-dimensional visualization of high-dimensional textual patterns for tourist reviews in the seven Northeastern Indian states. Sikkim and Mizoram signify a focused and localized cluster, with similar subject matter or common visitor experiences. Reviews for Assam and Meghalaya describe major differences in perceptions of landscapes, distinct attractions, and distinct visitor demographics. Arunachal and Nagaland showed visitors' perception of uniformity in common linguistic traits and tourism outlooks. Language-based interstate variations were also observed within the distinct groups based on perceptions and expectations of visitors or reviewers(Figure 8).

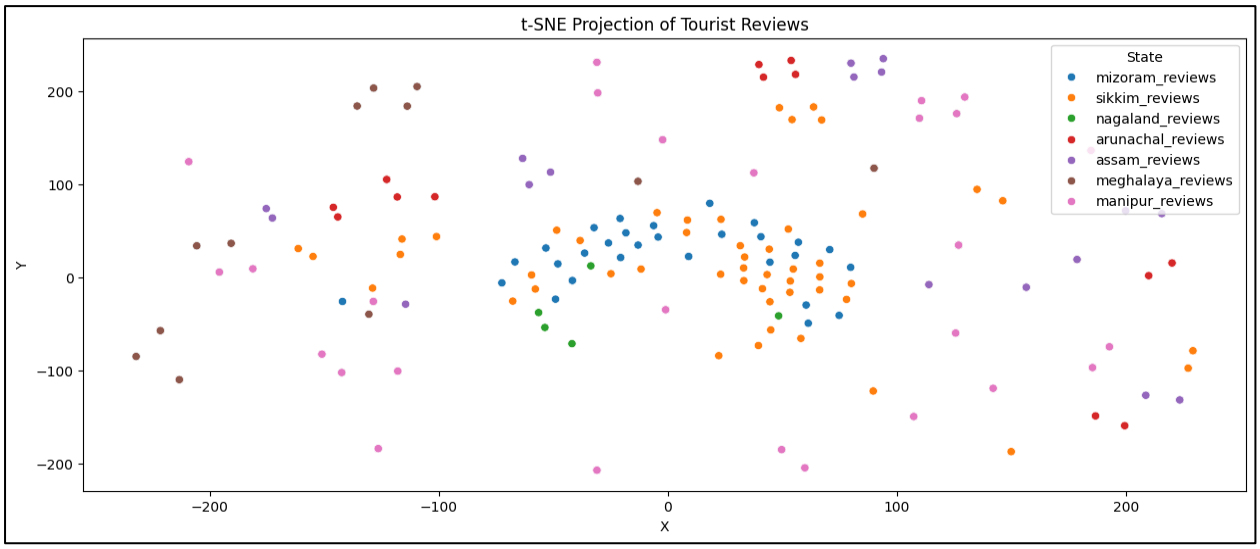


Fig. 8. t-SNE chart (Source: Author analysis)

5.8 Clustering using PCA

Three clusters were identified using the principal component analysis method. Cluster 1 included the states of Arunachal Pradesh, Assam, Meghalaya, and Sikkim (PCA component 1 range: -0.45 to -0.05), revealing similar review topics and similar tourist experience stories. Cluster 2 includes only Mizoram (PCA component 1 ≈ 0.0 and Component 2 > 0.6), with a unique review pattern. The third cluster exhibited Manipur and Nagaland (PCA Component 1 > 0.75), a mix of sentiment topic pairings. The distribution of clusters demonstrates three clear segmentations for efficient targeted marketing strategies(Figure 9).

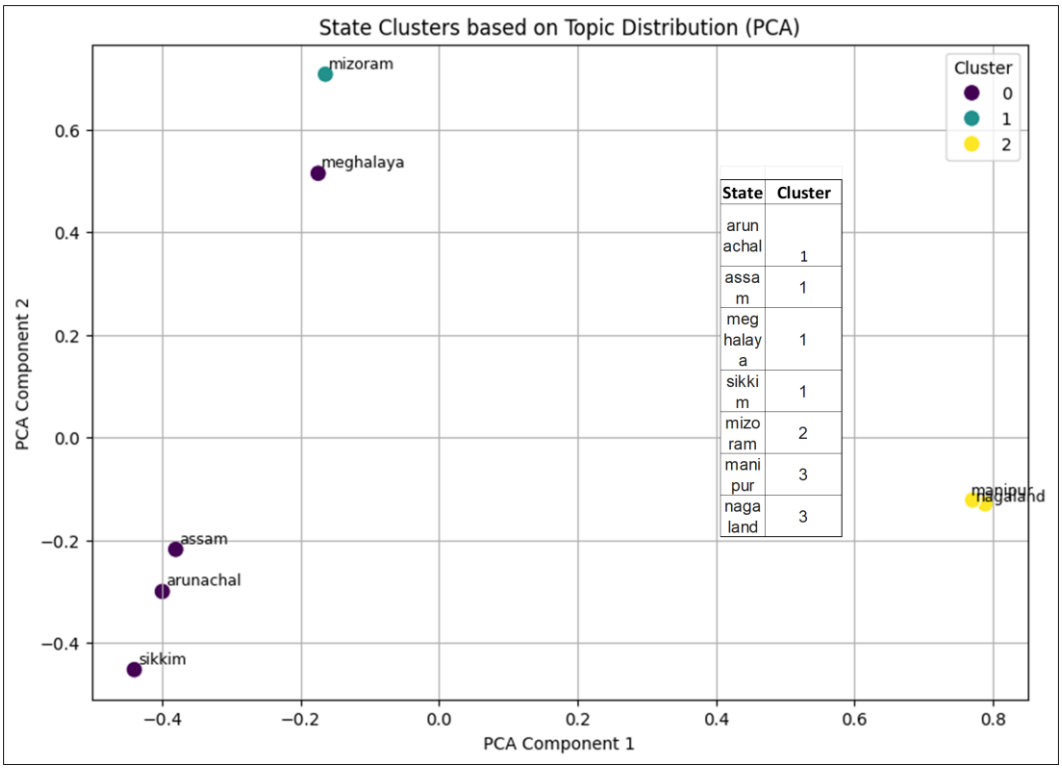


Fig. 9. t-SNE visualization of tourist reviews (Source: Author analysis)

5.9 Application of Sentiment Analysis Models to Regional Tourism Data

The purpose of sentiment analysis was to determine the emotional polarity of visitors' experience tone for the Indian northeastern states. This was ensured with systematic segmentation of reviews using lexicon-based methods (such as VADER and TextBlob) and machine learning approaches. Levels of personal opinions within the text were evaluated using subjectivity measurements, while the overall sentiments were determined using emotional polarity. Regional patterns (similarities and dissimilarities) of visitor satisfaction emerged for each state within the study. The present work follows a data-driven methodology for the systematic determination of experiences, emotions, and regional challenges from the statistically processed data collected from review websites, social media posts, and comments for the seven Indian northeastern states. It also identifies the relationship between sentiment scores, state-specific topics, and clusters for action-driven insights.

6. Major Findings and Conclusions

- Sentiment polarity and emotional tone of tourist reviews

Results from the analysis reveal significant variations in experience-based sentiment polarity and emotional tone of visitors. Sikkim and Arunachal Pradesh-based reviews highlight a positive emotional tone for natural beauty, adventurous activities, and cultural involvement. Some major challenges also emerged besides emotional tones. Positive comments for wildlife, tea farms, and traditional celebrations were accompanied by issues of accessibility, travel infrastructure, and transport management. Positive sentiments from emotional tone include joy, trust, and appreciation, while negative sentiments from emotional tone include frustration, disappointment, or doubt. The present work is suitable since it captures both polarity and complex sentiments in visitor-generated content, which improves interpretations of tourist perceptions and measures the quality of experience.

- Comparative sentiment trends among North-Eastern states

The polarity scores and visualization help in efficient segmentation for the states in North-Eastern India. Sentiment trends for each state were identified and compared for transparent segmentation across the Northeastern region. Comparative findings help to understand state-specific tones and the redesigning of customized market strategies for each of them.

- Insights into visitor satisfaction, challenges, and perceptions

Web-scraped reviews from northeastern states provide transparent insights into the visitor satisfaction, challenges, and general sentiments across the northeastern region. Visitor satisfaction relates to natural beauty, traditions, recreational opportunities, and the warmth of the residents in Sikkim, Arunachal Pradesh, and Mizoram. Challenges perceived in Assam, Manipur, and Nagaland were infrastructure, accessibility, seasonal crowding, and service quality. Cultural festivals, indigenous traditions, and local cuisine also influenced visitors' sentiments besides adventure and expedition opportunities in mountainous regions. These results support the need for customized tourism development plans for the improvement of visitor satisfaction levels and minimizing logistical and physical constraints across the region.

- Cluster formation using sentiment outcomes

Different clusters were formed for the Northeast Indian states. Visitors with positive sentiment for hiking and wildlife in Sikkim were part of the Eco-Adventure Cluster, whereas the Assam Cultural-Nature Hybrid Cluster represents a culture-rich but lack of infrastructure. Meghalaya's Nature-Cultural Cluster is identified by natural beauty and cultural attractions. Arunachal Pradesh's Adventure-Isolation Cluster promotes spirituality and remote adventure, whereas Manipur's Culture-Intimate Cluster has rich traditions but few amenities. Clusters signify the impact of emotions and experiences on regional tourism (see Appendix, Table 2).

7. Suggestions and Implications

- Summary of sentiment analysis outcomes and contributions

Sentiment analysis makes a significant contribution to determining emotion and perception across northeastern states. Positive emotions (natural beauty and hospitality) play a dominant role across all northeastern states despite existing challenges (infrastructure and accessibility). The efficiency of the research framework highlights its potential for tourism research at other destinations.

- Importance of sentiment analysis in regional and cultural contexts

Sentiment analysis of tourism reviews for northeast India highlights travelers' emotional reactions to regional culture, hospitality, and identity. Understanding sentiments can be linked to the influence of unique cultural characteristics in northeast India on tourist satisfaction. A more detailed examination of the states reveals that cultural differences like kindness, truthfulness, and local pride make tourists feel more positive about their visit. Sentiment analysis places emotions within regional and cultural contexts to provide opportunities for inclusive branding and destination management.

- Limitations and scope for future research in AI-driven tourism analytics

The proposed framework reflects the integration of sentiment analysis, clustering techniques, and text mining for high-quality insights into tourism in Northeast India. Present model interpretation is influenced by cultural context and emotional intensity, with some platform-specific biases, language variations, and limited rural representation. Future analysis may use inputs such as videos, images, geotags, and complex AI models. Longitudinal studies can be combined with monitoring technologies for advanced behavioral analysis for real-time tourism intelligence systems. All these developments can ensure a more sustainable and culturally sensitive development in evolving regional destinations.

Acknowledgement
This research was not funded by any grant.

Conflicts of Interest
The authors declare no conflicts of interest.

References

[1] Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature. https://link.springer.com/book/10.1007/978-3-031-02145-9
[2] Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems, 53(4), 675-679. https://doi.org/10.1016/j.dss.2012.05.022
[3] Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing, 40(1), 75-87. https://doi.org/10.1016/j.ijresmar.2022.05.005
[4] Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002
[5] Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. Procedia Computer Science, 152, 341-348. https://doi.org/10.1016/j.procs.2019.05.008

[6] Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. Decision support systems, 53(4), 704-711. https://doi.org/10.1016/j.dss.2012.05.023
[7] Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications, 69, 214-224. https://doi.org/10.1016/j.eswa.2016.10.043
[8] Poria, S., Cambria, E., Winterstein, G., & Huang, G. B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. Knowledge-Based Systems, 69, 45-63. https://doi.org/10.1016/j.knosys.2014.05.005
[9] Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. Procedia Computer Science, 87, 44-49. https://doi.org/10.1016/j.procs.2016.05.124
[10] Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Systems, 52, 201-213. https://doi.org/10.1016/j.knosys.2013.08.011
[11] Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. Telematics and Informatics, 35(1), 136-147. https://doi.org/10.1016/j.tele.2017.10.006
[12] Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. Artificial intelligence in medicine, 64(1), 17-27. https://doi.org/10.1016/j.artmed.2015.03.006
[13] Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision support systems, 55(4), 919-926. https://doi.org/10.1016/j.dss.2012.12.028
[14] Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of travel research, 58(2), 175-191. https://doi.org/10.1177/0047287517747753
[15] Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016
[16] Farhadloo, M., & Rolland, E. (2016). Fundamentals of sentiment analysis and its applications. In Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence (pp. 1-24). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-319-30319-2_1
[17] Chaturvedi, I., Cambria, E., Welsch, R. E., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. Information Fusion, 44, 65-77. https://doi.org/10.1016/j.inffus.2017.12.006
[18] Clavel, C., & Callejas, Z. (2015). Sentiment analysis: from opinion mining to human-agent interaction. IEEE Transactions on affective computing, 7(1), 74-93. https://doi.org/10.1109/taffc.2015.2444846
[19] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In A practical guide to sentiment analysis (pp. 1-10). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-55394-8_1
[20] Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing, 40(1), 75-87. https://doi.org/10.1016/j.ijresmar.2022.05.005
[21] Hu, Y., Deng, C., & Zhou, Z. (2019). A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. Annals of the American Association of Geographers, 109(4), 1052-1073. https://doi.org/10.1080/24694452.2018.1535886
[22] Reyes-Menendez, A., Saura, J. R., & Alvarez-Alonso, C. (2018). Understanding# WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach. International journal of environmental research and public health, 15(11), 2537. https://doi.org/10.3390/ijerph15112537
[23] Mehraliyev, F., Chan, I. C. C., & Kirilenko, A. P. (2022). Sentiment analysis in hospitality and tourism: a thematic and methodological review. International Journal of Contemporary Hospitality Management, 34(1), 46-77. https://doi.org/10.1108/ijchm-02-2021-0132
[24] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011
[25] Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. Expert systems with applications, 40(10), 4065-4074. https://doi.org/10.1016/j.eswa.2013.01.001
[26] Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., ... & Kwak, K. S. (2019). Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowledge-Based Systems, 174, 27-42. https://doi.org/10.1016/j.knosys.2019.02.033
[27] Lau, R. Y., Li, C., & Liao, S. S. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. Decision Support Systems, 65, 80-94. https://doi.org/10.1016/j.dss.2014.05.005
[28] Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Computing Surveys, 55(13s), 1-38. https://doi.org/10.1145/3586075

[29] Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77, 236-246. https://doi.org/10.1016/j.eswa.2017.02.002
[30] Da Silva, N. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision support systems, 66, 170-179. https://doi.org/10.1016/j.dss.2014.07.003
[31] Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. Computers in Human Behavior, 56, 179-191. https://doi.org/10.1016/j.chb.2015.11.040
[32] Sánchez-Núñez, P., Cobo, M. J., De Las Heras-Pedrosa, C., Peláez, J. I., & Herrera-Viedma, E. (2020). Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis. IEEE Access, 8, 134563-134576. https://doi.org/10.1109/access.2020.3009482
[33] Puschmann, C., & Powell, A. (2018). Turning words into consumer preferences: How sentiment analysis is framed in research and the news media. Social Media+ Society, 4(3), 2056305118797724. https://doi.org/10.1177/2056305118797724
[34] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in human behavior, 31, 527-541. https://doi.org/10.1016/j.chb.2013.05.024
[35] Loureiro, M. L., & Alló, M. (2020). Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the UK and Spain. Energy Policy, 143, 111490. https://doi.org/10.1016/j.enpol.2020.111490
[36] Mehraliyev, F., Chan, I. C. C., & Kirilenko, A. P. (2022). Sentiment analysis in hospitality and tourism: a thematic and methodological review. International Journal of Contemporary Hospitality Management, 34(1), 46-77. https://doi.org/10.1108/ijchm-02-2021-0132
[37] Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. Expert Systems with Applications, 167, 114324. https://doi.org/10.1016/j.eswa.2020.114324
[38] Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. Journal of Travel Research, 57(8), 1012-1025. https://doi.org/10.1177/0047287517729757
[39] Ali, T., Marc, B., Omar, B., Soulaimane, K., & Larbi, S. (2021). Exploring destination's negative e-reputation using aspect based sentiment analysis approach: Case of Marrakech destination on TripAdvisor. Tourism Management Perspectives, 40, 100892. https://doi.org/10.1016/j.tmp.2021.100892
[40] Kim, K., Park, O. J., Yun, S., & Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. Technological Forecasting and Social Change, 123, 362-369. https://doi.org/10.1016/j.techfore.2017.01.001
[41] Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., ... & Tang, J. (2022). Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. Heliyon, 8(10). https://doi.org/10.1016/j.heliyon.2022.e10894
[42] Obembe, D., Kolade, O., Obembe, F., Owoseni, A., & Mafimisebi, O. (2021). Covid-19 and the tourism industry: An early stage sentiment analysis of the impact of social media and stakeholder communication. International Journal of Information Management Data Insights, 1(2), 100040. https://doi.org/10.1016/j.jjimei.2021.100040
[43] Tao, Y., Zhang, F., Shi, C., & Chen, Y. (2019). Social media data-based sentiment analysis of tourists' air quality perceptions. Sustainability, 11(18), 5070. https://doi.org/10.3390/su11185070
[44] Philander, K., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. International Journal of Hospitality Management, 55(2016), 16-24. https://doi.org/10.1016/j.ijhm.2016.02.001
[45] Mirzaalian, F., & Halpenny, E. (2021). Exploring destination loyalty: Application of social media analytics in a nature-based tourism setting. Journal of Destination Marketing & Management, 20, 100598. https://doi.org/10.1016/j.jdmm.2021.100598
[46] Barbierato, E., Bernetti, I., & Capecchi, I. (2022). Analyzing TripAdvisor reviews of wine tours: an approach based on text mining and sentiment analysis. International Journal of Wine Business Research, 34(2), 212-236. https://doi.org/10.1108/ijwbr-04-2021-0025
[47] Ramanathan, V., & Meyyappan, T. (2019, January). Twitter text mining for sentiment analysis on people's feedback about Oman tourism. In 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC) (pp. 1-5). IEEE. https://doi.org/10.1109/icbdsc.2019.8645596
[48] Ali, F., Kwak, D., Khan, P., Islam, S. R., Kim, K. H., & Kwak, K. S. (2017). Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transportation Research Part C: Emerging Technologies, 77, 33-48. https://doi.org/10.1016/j.trc.2017.01.014
[49] Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2014, January). Sentiment analysis: Extracting decision-relevant knowledge from UGC. In Information and Communication Technologies in Tourism 2014: Proceedings of

the International Conference in Dublin, Ireland, January 21-24, 2014 (pp. 253-265). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-03973-2_19

[50] Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism management, 58, 51-65. https://doi.org/10.1016/j.tourman.2016.10.001

[51] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

[52] Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Emotion Measurement (2nd ed., pp. 323–379). Elsevier. https://doi.org/10.1016/B978-0-12-818554-8.00013-4

[53] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89. https://doi.org/10.1145/2436256.2436274

[54] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

[55] Aggarwal, C. C., & Zhai, C. (2012). An introduction to text mining. In Mining text data (pp. 1-10). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-3223-4_1

[56] Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. International journal of advanced computer science and applications, 7(11), 414-418. https://doi.org/10.14569/ijacsa.2016.071153

[57] Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. Information processing & management, 43(5), 1216-1247. https://doi.org/10.1016/j.ipm.2006.11.011

[58] Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). Using text mining techniques for extracting information from research articles. In Intelligent natural language processing: Trends and Applications (pp. 373-397). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-67056-0_18

[59] Chowdhary, K. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649. https://doi.org/10.1007/978-81-322-3972-7_19

[60] Deng, L., & Liu, Y. (Eds.). (2018). Deep learning in natural language processing. Springer. https://doi.org/10.1007/978-981-10-5209-5

[61] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266. https://doi.org/10.1126/science.aaa8685

[62] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications, 82(3), 3713-3744. https://doi.org/10.1007/s11042-022-13428-4

[63] Meystre, S., & Haug, P. J. (2006). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. Journal of biomedical informatics, 39(6), 589-599. https://doi.org/10.1016/j.jbi.2005.11.004

[64] Galati, F., & Bigliardi, B. (2019). Industry 4.0: Emerging themes and future research avenues using a text mining approach. Computers in industry, 109, 100-113. https://doi.org/10.1016/j.compind.2019.04.018

[65] Banjar, A., Ahmed, Z., Daud, A., Abbasi, R. A., & Dawood, H. (2021). Aspect-Based Sentiment Analysis for Polarity Estimation of Customer Reviews on Twitter. Computers, Materials & Continua, 67(2). https://doi.org/10.32604/cmc.2021.014226

[66] Mowlaei, M. E., Abadeh, M. S., & Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. Expert Systems with Applications, 148, 113234. https://doi.org/10.1016/j.eswa.2020.113234

[67] Venugopalan, M., & Gupta, D. (2022). An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. Knowledge-based systems, 246, 108668. https://doi.org/10.1016/j.knosys.2022.108668

[68] Ali, T., Marc, B., Omar, B., Soulaimane, K., & Larbi, S. (2021). Exploring destination's negative e-reputation using aspect based sentiment analysis approach: Case of Marrakech destination on TripAdvisor. Tourism Management Perspectives, 40, 100892. https://doi.org/10.1016/j.tmp.2021.100892

[69] Mehra, P. (2023). Unexpected surprise: Emotion analysis and aspect based sentiment analysis (ABSA) of user generated comments to study behavioral intentions of tourists. Tourism Management Perspectives, 45, 101063. https://doi.org/10.1016/j.tmp.2022.101063

[70] Nazir, A., Rao, Y., Wu, L., & Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. IEEE Transactions on Affective Computing, 13(2), 845-863. https://doi: 10.1109/TAFFC.2020.2970399

[71] He, J., Wumaier, A., Kadeer, Z., Sun, W., Xin, X., & Zheng, L. (2022). A local and global context focus multilingual learning model for aspect-based sentiment analysis. IEEE Access, 10, 84135-84146. https://doi.org/10.1109/access.2022.3197218
[72] Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. Ieee Access, 8, 181074-181090. https://doi.org/10.1109/access.2020.3027350
[73] Sharma, N. A., Ali, A. S., & Kabir, M. A. (2025). A review of sentiment analysis: tasks, applications, and deep learning techniques. International journal of data science and analytics, 19(3), 351-388. https://doi.org/10.1007/s41060-024-00594-x
[74] Alahmadi, K., Alharbi, S., Chen, J., & Wang, X. (2025). Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. Social Network Analysis and Mining, 15(1), 1-28. https://doi.org/10.1007/s13278-025-01461-8
[75] Thompson, R. A. (1994). Emotion regulation: A theme in search of definition. Monographs of the society for research in child development, 25-52. https://doi.org/10.1111/j.1540-5834.1994.tb01276.x
[76] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic markets, 31(3), 685-695. https://doi.org/10.1007/s12525-021-00475-2
[77] Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018, August). Comparative study between traditional machine learning and deep learning approaches for text classification. In Proceedings of the ACM Symposium on Document Engineering 2018 (pp. 1-11). https://doi.org/10.1145/3209280.3209526
[78] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Molecular diversity, 25(3), 1315-1360.
[79] Ahmad, J., Farman, H., & Jan, Z. (2018). Deep learning methods and applications. In Deep learning: convergence to big data analytics (pp. 31-42). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-3459-7_3
[80] Chauhan, N. K., & Singh, K. (2018, September). A review on conventional machine learning vs deep learning. In 2018 International conference on computing, power and communication technologies (GUCON) (pp. 347-352). IEEE. https://doi.org/10.1109/gucon.2018.8675097
[81] Rana, M., & Bhushan, M. (2023). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications, 82(17), 26731-26769. https://doi.org/10.1007/s11042-022-14305-w
[82] Kocher, G., & Kumar, G. (2021). Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges. Soft Computing, 25(15), 9731-9763. https://doi.org/10.1007/s00500-021-05893-0
[83] Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug discovery today, 22(11), 1680-1685. https://doi.org/10.1016/j.drudis.2017.08.010
[84] Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for naive Bayes text classifiers. Knowledge-Based Systems, 100, 137-144. https://doi.org/10.1016/j.knosys.2016.02.017
[85] Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128. https://doi.org/10.1016/j.protcy.2012.05.017
[86] Christmann, A., & Steinwart, I. (2008). Support vector machines. https://doi.org/10.1007/978-0-387-77242-4_9
[87] Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. Artificial Intelligence Review, 52(2), 803-855. https://doi.org/10.1007/s10462-018-9614-6
[88] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222. https://doi.org/10.1080/01431160412331269698
[89] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics, 10(1), 213. https://doi.org/10.1186/1471-2105-10-213
[90] Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resources Management, 31(9), 2761-2775. https://doi.org/10.1007/s11269-017-1660-3
[91] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232. https://doi.org/10.1109/tnnls.2016.2582924
[92] Sahoo, B. B., Jha, R., Singh, A., & Kumar, D. (2019). Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophysica, 67(5), 1471-1481. https://doi.org/10.1007/s11600-019-00330-1

[93] Muzaffar, S., & Afshari, A. (2019). Short-term load forecasts using LSTM networks. Energy Procedia, 158, 2922-2927.https://doi.org/10.1016/j.egypro.2019.01.952
[94] Cortez, B., Carrera, B., Kim, Y. J., & Jung, J. Y. (2018). An architecture for emergency event prediction using LSTM recurrent neural networks. Expert Systems with Applications, 97, 315-324. https://doi.org/10.1016/j.eswa.2017.12.037
[95] Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., & El Asri, B. (2021). BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update, 1, 100042. https://doi.org/10.1016/j.cmpbup.2021.100042
[96] Lin, S. Y., Kung, Y. C., & Leu, F. Y. (2022). Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis. Information Processing & Management, 59(2), 102872. https://doi.org/10.1016/j.ipm.2022.102872
[97] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.
[98] Kim, J., Kim, H. S., & Choi, S. Y. (2023). Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: A FinBERT transformer model and LSTM. Axioms, 12(9), 835. https://doi.org/10.3390/axioms12090835
[99] Github,2025 , Available at https://github.com/sandeepbitmba/tourism-data.

Appendices

Table 2
State-wise Analysis and hypothesis

Si. No.	State	Type of Analysis	Interpretation	Hypothesis
1	Sikkim	Sentiment + Topic Modeling + Clustering	High positive sentiment; dominant themes: scenic beauty, cleanliness, trekking; falls in **eco-adventure** cluster.	H1a: Sikkim's strong eco-adventure appeal drives consistently high positive sentiment.
2	Assam	Sentiment + Topic Modeling + Clustering	Mixed sentiment; themes: wildlife, tea estates, cultural festivals; clusters into a **cultural-nature hybrid**.	H1b: The sentiment equilibrium in Assam is impacted by its cultural richness despite the challenges posed by its infrastructure.
3	Meghalaya	Sentiment + Topic Modeling + Clustering	Balanced sentiment; themes: waterfalls, caves, indigenous culture; in a **nature-cultural** cluster.	H1c: Seasonal crowding moderates the otherwise high positive sentiment for Meghalaya.
4	Arunachal Pradesh	Sentiment + Topic Modeling + Clustering	High positive sentiment; themes: adventure, remoteness, monasteries; in **adventure-isolation** cluster.	H1d: Remote adventure experiences overcome accessibility problems while shaping sentiment about Arunachal Pradesh.
5	Manipur	Sentiment + Topic Modeling + Clustering	Mixed sentiment; themes: an exploration of cultural immersion, dance festivals, and local crafts within a **closely-knit** cultural cluster.	H1e: Limited infrastructure reduces sentiment despite rich cultural offerings in Manipur.
6	Mizoram	Sentiment + Topic Modeling + Clustering	Positive sentiment; themes: hospitality, cuisine, beautiful landscapes; within the leisure and hospitality sector.	H1f: Favorable views on hospitality compensate for the difficulties posed by a low number of tourists in Mizoram.
7	Nagaland	Sentiment + Topic Modeling + Clustering	Moderate sentiment; themes: Hornbill Festival, tribal heritage, remote experiences; in **festival-cultural** cluster.	H1g: Seasonal festival events significantly influence sentiment trends in Nagaland.

Source: Calculated by author