INSIGHTS INTO VISITOR PERCEPTIONS OF KAZAKHSTAN'S TOURISM SECTOR: SENTIMENT ANALYSIS AND TOPIC MODELING OF TRIPADVISOR REVIEWS

By

Asset Kabdula

Submitted to
Central European University
Department of Network and Data Science

In partial fulfillment of the requirements for the degree of
Master in Social Data Science

Supervisor: Prof. Marton Posfai

Vienna, Austria

2025

Author's Declaration

I, the undersigned, **Asset Kabdula**, candidate for the MS degree in Social Data Science declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 22 May 2025

Asset Kabdula

Signature

Copyright Notice

[1] Icon by Font Awesome: https://fontawesome.com/

Abstract

Kazakhstan's tourism industry has experienced rapid growth in recent years, yet academic research on tourists' perceptions and satisfaction in this context remains limited. This study leverages user-generated content from TripAdvisor to provide a comprehensive sentiment and topic analysis of visitor experiences across Kazakhstan's attractions from 2011 to 2024. Employing advanced machine learning and natural language processing techniques including BERT-based sentiment classification, Non-negative Matrix Factorization (NMF) for topic modeling, and zero-shot topic classification over 23,000 multilingual reviews covering 1,001 attractions were systematically analyzed.

The results reveal a strong positive correlation between TripAdvisor review activity and official international tourist arrivals, supporting the use of digital trace data as a proxy for real-world tourism trends. Sentiment analysis indicates an overall upward trend in positive sentiment among tourists, particularly after 2020, coinciding with the State Program for the Development of the Tourism Industry (2019–2025). Topic modeling identifies seven key themes such as transportation, tourism services, culture and history, nature and hiking, religious and spiritual sites, urban parks, and food and leisure with transportation-related reviews consistently associated with lower sentiment, highlighting persistent challenges in infrastructure. Conversely, nature and hiking topics exhibit the highest sentiment scores, emphasizing Kazakhstan's natural attractions as a key strength.

The study further finds no local bias in online review data and finds no significant difference in sentiment between reviews submitted by Kazakhstanis and those provided by foreign visitors. Additionally, there is a statistically significant seasonality in sentiment, with peaks in November and troughs in March. These findings offer actionable insights for policymakers, destination marketers, and tourism authorities, underscoring the value of digital review analytics in enhancing tourism strategies and service quality. By filling a significant gap in the literature, this research provides a robust, data-driven understanding of tourist perceptions and evolving trends in Kazakhstan's tourism sector.

Acknowledgements

I would like to thank Professor Marton Posfai from Central European University for his unwavering support, encouragement, and invaluable guidance throughout the course of my research. His constructive feedback and insightful suggestions have greatly contributed to the quality and completion of this thesis. I am deeply grateful for his mentorship and for inspiring me to pursue academic excellence. I am profoundly grateful to my family for always believing in me and supporting me every step of the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Kazakhstan's tourism industry has experienced significant growth in recent years, driven by its diverse landscapes, rich cultural heritage, and government initiatives aimed at attracting international visitors. The country offers a unique blend of natural wonders, including the Charyn Canyon, Altai Mountains, and the Caspian coastline, alongside historic sites such as the Mausoleum of Khoja Ahmed Yasawi and the Silk Road cities of Turkestan and Shymkent [1]. These attractions make Kazakhstan a compelling destination for both domestic and international tourists.

In the digital age, tourist experiences are increasingly shared online through platforms such as TripAdvisor, which influence potential visitors' perceptions and decision-making processes. Sentiment analysis, a subfield of natural language processing (NLP), has emerged as a valuable tool for extracting insights from user-generated content on these platforms. Previous studies have successfully applied sentiment analysis techniques to tourism research, helping stakeholders understand tourist satisfaction and identify areas for improvement [2].

Existing research has largely focused on well-established global tourism destinations, such as European and Southeast Asian countries, analyzing sentiment trends in hospitality services, transportation, and destination attractiveness [3]. However, there has been limited academic focus on sentiment analysis in Kazakhstan's tourism sector, leaving a gap in understanding how tourists perceive and experience the country.

Despite the growing importance of digital reviews in shaping tourism experiences, there is a noticeable gap in sentiment analysis research focusing on Kazakhstan. While studies have explored sentiment trends in other destinations, limited research has leveraged tourist-generated content to assess satisfaction in Kazakhstan's tourism sector. Understanding tourist sentiment can provide valuable insights for policymakers, businesses, and marketers to enhance services and promote the country as a desirable travel destination [4].

This research aims to fill this gap by analyzing sentiment trends in tourist reviews of Kazakhstan, identifying key themes, and uncovering factors that influence positive and negative perceptions. By leveraging machine learning and NLP techniques, this study contributes to the growing field of sentiment analysis in tourism research while providing practical recommendations for stakeholders in Kazakhstan's tourism industry.

The central research question for this study is: **What key themes emerge from sentiment analysis and topic modeling of Kazakhstan's tourism sector based on digital reviews?**

This study analyzes user-generated content from TripAdvisor to examine tourist sentiment toward Kazakhstan's attractions, services, and overall experience. The research employs machine learning and deep learning models, particularly BERT-based sentiment analysis, to assign positive sentiment scores to reviews. Additionally, natural language processing tools such as Non-negative Matrix Factorization (NMF) and Term Frequency-Inverse Document Frequency (TF-IDF) are used to identify recurring themes such as transportation, culture and history, nature and hiking, food and leisure, religious and spiritual sites, urban parks and tourism services.

The methodology involves data collection through web scraping from TripAdvisor platform followed by data preprocessing, including tokenization and English stop word removal, and finally sentiment scoring and topic modeling. The results are visualized using sentiment trend graphs to identify geographical variations in tourist satisfaction. By analyzing sentiment over time, this study also examines whether Kazakhstan's tourism sector shows signs of improvement in response to visitor feedback.

The study hypothesizes the following:

- **H1** Tourist sentiment in digital reviews demonstrates a positive and increasing trend over time in Kazakhstan, with the most favorable perceptions consistently associated with the country's natural attractions.

- **H2** Reviews on TripAdvisor written by Kazakhstanis do not exhibit significant bias toward their own country's attractions compared to reviews by international visitors.

- **H3** Positive sentiment scores in digital reviews for Kazakhstan's attractions exhibit significant seasonal variation, with the average sentiment score systematically changing by month across the study period.

By providing a data-driven approach to understanding Kazakhstan's tourism industry, this research aims to offer actionable insights for destination marketers, tourism authorities, and local businesses to enhance the overall visitor experience. The findings are relevant for developing targeted marketing strategies and improving service quality based on tourists' real experiences and expectations.

Chapter 2

Literature Review

Sentiment analysis has been widely applied in tourism research to assess visitor satisfaction, identify trends, and guide decision-making for industry stakeholders. Recent studies have demonstrated the effectiveness of NLP and machine learning techniques in extracting meaningful insights from online reviews. For instance, Charfaoui and Mussard (2024) [1] utilized sentiment analysis to understand the emotional tone of tourist feedback, highlighting key aspects that contribute to visitor satisfaction. Similarly, Enache (2020) [4] explored sentiment trends in tourism-related content, showing how sentiment analysis helps destinations tailor their marketing strategies.

Several studies have explored sentiment analysis across different tourism sectors, such as hospitality, transportation, and destination marketing. George and Ramos (2024) [5] applied sentiment analysis to wellness tourism destinations, revealing how visitor experiences and emotions influence brand perception. Nawawi et al. (2024) [2] employed aspect-based sentiment analysis with zero-shot learning techniques to assess hospitality services, demonstrating the role of advanced NLP models in improving service quality.

In the domain of transportation, Pineda-Jaramillo et al. (2023) [6] investigated tourist satisfaction regarding transport systems in Mount Etna, emphasizing how transportation infrastructure affects overall visitor experiences. Valdivia, Luzón, and Herrera (2017) [3] focused on sentiment analysis in TripAdvisor reviews, showcasing the role of online platforms in shaping destination reputation.

While sentiment analysis has been extensively applied to popular tourist destinations, there remains a lack of research on Kazakhstan's tourism sector. Studies focusing on European and Southeast Asian countries have provided valuable insights into sentiment trends, but their findings may not be directly applicable to Kazakhstan due to differences in cultural, economic, and infrastructural factors. Moreover, research incorporating Kazakh tourism data using advanced machine learning models is limited, highlighting the need for a comprehensive study in this context.

This study aims to bridge the gap by conducting sentiment analysis on digital reviews of Kazakhstan's tourism sector. By leveraging BERT-based sentiment classification and topic modeling, this research provides a nuanced understanding of tourist perceptions, key themes, and factors influencing visitor satisfaction. The findings contribute to the literature by offering data-driven insights specific to Kazakhstan, enabling policymakers and businesses to make informed decisions to enhance the country's tourism industry.

This study contributes to the literature by applying sentiment analysis techniques to Kazakhstan's tourism sector, identifying key themes in tourist reviews. It provides insights into how different factors such as seasonality, service quality, and regional attractions impact sentiment trends. By employing advanced NLP-based models, such as BERT, this research offers a data-driven approach to understanding tourist satisfaction and enhancing tourism development strategies in Kazakhstan.

Chapter 3

Methodology

This research adopts a multi-stage methodology that integrates advanced natural language processing and machine learning techniques to analyze tourist perceptions of Kazakhstan. The methodological workflow consists of several key steps: (1) large-scale data collection and cleaning of TripAdvisor reviews, (2) multilingual sentiment analysis using transformer-based models, (3) translation and preprocessing of text data for consistent topic modeling, (4) identification of core topics using Non-negative Matrix Factorization (NMF), (5) topic-specific sentiment evaluation with zero-shot classification, and (6) statistical analysis to assess local bias and seasonality in visitor sentiment.

3.1 Data Collection and Preparation

The primary source of data for this study consists of tourist reviews obtained from TripAdvisor, collected through web scraping techniques. Specifically, reviews were gathered from the 1,000 most popular tourist attractions in Kazakhstan listed on the platform, resulting in a dataset covering 1,001 attractions. The web scraping process involved automated retrieval of both review content and associated metadata using Python libraries such as Requests, BeautifulSoup, and Selenium for dynamic content rendering and interaction with JavaScript elements. Web scraping posed several challenges due to TripAdvisor's robust anti-bot mechanisms, including CAPTCHA verification and automated bot detection systems. Despite these obstacles, comprehensive scraping was achieved, yielding a detailed dataset containing 825 variables.

Following data collection, rigorous data cleaning was conducted to ensure the dataset's relevance and quality. Initially, only attractions with at least one review were retained. Subsequently, irrelevant variables were removed, retaining only those directly pertinent to the research objectives. The refined dataset comprises 18 variables categorized into:

- **Attraction information:** id, placeInfo/name, placeInfo/locationString, placeInfo/latitude, placeInfo/longitude, placeInfo/numberOfReviews, placeInfo/rating, placeInfo/website, locationId.

- **Review details:** rating, title, text, publishedDate, travelDate, tripType, helpfulVotes.
- **User metadata:** lang, user/userLocation/name.

This final dataset contains 23,296 individual reviews in 26 languages, covering the period from 2011 to early 2025. Reviews from 2025 were excluded from subsequent analyses, as data collection occurred at the beginning of the year, thus potentially misrepresenting the full year's review activity.

To supplement this dataset and validate its representativeness, official tourism statistics from the Bureau of National Statistics of Kazakhstan were acquired. Specifically, data on incoming tourist arrivals from 2011 to 2022 were collected. This secondary dataset enabled the assessment of the correlation between review frequency on TripAdvisor and actual tourist inflow, examining whether online review activity reliably reflects real-world tourism trends.

3.2 Correlation Between TripAdvisor Reviews and Tourist Arrivals

To assess whether user-generated review activity on TripAdvisor reflects actual tourism flows, a correlation analysis was conducted between the annual number of TripAdvisor reviews and official statistics on international tourist arrivals to Kazakhstan. Annual counts of TripAdvisor reviews were derived from the collected dataset, while official arrival data were obtained from the Bureau of National Statistics of Kazakhstan for the period from 2010 to 2022 [7]. Pearson's correlation coefficient was used to quantify the linear relationship between these two variables. The Pearson correlation coefficient (r) measures the strength and direction of a linear relationship between two continuous variables, and is calculated as the covariance of the variables divided by the product of their standard deviations. The coefficient ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship, and values around 0 suggest no linear association [8]. The statistical significance of the correlation was tested to determine whether increased online review activity is reliably associated with higher numbers of tourist arrivals. Through this analysis, the potential of digital trace data to serve as a proxy for real-world tourism trends in Kazakhstan was evaluated.

3.3 Sentiment Analysis (H1)

Sentiment analysis for this study was performed using the **cardiffnlp/twitter-xlm-roberta-base-sentiment** model, a transformer-based architecture available through the Hugging Face platform [9]. This model is built upon the XLM-RoBERTa-base foundation, a multilingual variant of RoBERTa that has been pre-trained on a large corpus of text in multiple languages.

The cardiffnlp/twitter-xlm-roberta-base-sentiment model is specifically fine-tuned on multilingual Twitter datasets for sentiment classification tasks, which makes it especially well-suited for the present analysis of diverse, informal, and multilingual review texts.

The model is capable of assigning sentiment labels to positive, neutral, or negative, and is optimized to handle informal language, abbreviations, and colloquialisms often found in online reviews. Its robust performance on social media and review data ensures consistent and accurate sentiment detection across a wide range of languages and writing styles present in the dataset. The computational analysis was performed locally using an Apple MacBook Pro with an M3 Pro chip, leveraging GPU acceleration for efficient and rapid inference.

The sentiment analysis model provides probabilistic outputs across three sentiment categories: positive, neutral, and negative, whose scores sum to one. For analytical purposes, the positive sentiment probability was used as the primary indicator of the review's overall tone. This decision is justified based on its direct relevance to tourism marketing, robustness to neutral textual noise, and ease of interpretability in statistical analyses. Sensitivity analyses employing the negative sentiment probability were also conducted, confirming the robustness of this approach.

The choice of a transformer-based model was driven by its demonstrated superior performance compared to traditional statistical techniques such as TF-IDF-based classifiers [10]. Transformer architectures possess sophisticated attention mechanisms, allowing them to capture intricate contextual and semantic relationships within texts, including word order, polysemy, and long-range dependencies. Consequently, transformer-based models achieve significantly higher performance across key metrics (accuracy, precision, recall, and F1 score), albeit at the cost of increased computational resources [11].

Another limitation of transformer-based models, particularly those hosted on platforms such as Hugging Face, is their susceptibility to biases encoded during training. These biases arise from imbalanced training datasets, which often reflect societal prejudices, stereotypes, or skewed perspectives present in the original data sources [12] [13]. Such biases can manifest in sentiment analysis tasks as unfairly skewed sentiment predictions toward particular demographics, cultures, or topics, thereby perpetuating harmful stereotypes or providing misleading insights [14]. Addressing this issue typically requires careful dataset curation, bias evaluation, and ongoing efforts to fine-tune models with balanced and diverse training data. While Hugging Face offers tools and metrics to detect and mitigate bias, practitioners must remain vigilant in interpreting model outputs critically and transparently acknowledging potential biases in their analyses.

3.4 Topic Modeling

Topic modeling was used to identify key themes discussed in tourist reviews. This process began with translating all reviews into English, followed by text preprocessing and feature extraction. The following sections describe each step in detail.

3.4.1 Translation for Topic Modeling

To enable advanced analyses such as topic modeling and topic-specific sentiment evaluation, all non-English reviews in the dataset were translated into English prior to processing. Translation was carried out using the **facebook/nllb-200-distilled-600M** transformer model, a cutting-edge neural machine translation system developed as part of Meta's No Language Left Behind (NLLB) initiative [15]. This model is specifically designed to support 200 languages, with a focus on delivering high-quality translations for both high-resource and low-resource languages making it especially suitable for datasets containing linguistic diversity such as that of Kazakhstan's tourism sector.

The **facebook/nllb-200-distilled-600M** model is a distilled version of the original NLLB-200, providing an optimal balance between translation accuracy and computational efficiency [15]. Leveraging a massive multilingual training corpus, the model employs deep transformer-based architectures to capture nuanced linguistic structures and deliver contextually appropriate translations. Its performance on low-resource languages and domain-specific content has been shown to outperform previous state-of-the-art systems, reducing translation errors and preserving the sentiment and intent of the source text. The translation process was conducted using GPU acceleration on an Apple MacBook Pro equipped with an M3 Pro chip, ensuring computational efficiency and accuracy.

3.4.2 Text Pre-processing

The textual content of the reviews underwent systematic pre-processing to prepare the data for topic modeling. Pre-processing is essential in natural language processing (NLP) to normalize textual data, reduce noise, and enhance model accuracy. The following pre-processing pipeline was implemented:

- **Normalization:** Text was converted to lowercase.
- **Cleaning:** Digits, punctuation, special characters, and extraneous whitespace were removed.
- **Tokenization:** Text was split into individual tokens (words).
- **Stop-word Removal:** Common English stop words that do not convey significant meaning were eliminated.
- **Lemmatization:** Tokens were lemmatized using Part-of-Speech (POS) tagging, which involves assigning grammatical tags (e.g., noun, verb, adjective) to each word. Lemmatization reduces words to their base forms (e.g., ”changing,” ”changed,” and ”change” to ”change”), ensuring semantically similar words are analyzed collectively.

This comprehensive text pre-processing approach facilitated accurate and meaningful feature extraction, which is critical for subsequent analytical tasks.

3.4.3 Word Frequency Using TF-IDF

For machine learning models to process text data, words must be represented in a numerical format, a process known as vectorization. One widely used method in natural language processing (NLP) is Term Frequency–Inverse Document Frequency (TF-IDF), which quantifies the importance of a term within a specific document relative to a collection of documents (the corpus).

The TF-IDF score for term t in document d is computed as follows:

1. **Term Frequency (TF):**

$$\text{TF}(t,d) = \frac{\text{Number of occurrences of } t \text{ in } d}{\text{Total number of terms in } d}$$

2. **Inverse Document Frequency (IDF):**

$$\text{IDF}(t) = \log\left(\frac{D}{|\{d \in D : t \in d\}|}\right)$$

   where D is the total number of documents in the corpus, and $|\{d \in D : t \in d\}|$ is the number of documents containing the term t.

3. **TF-IDF Score:**

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

TF-IDF assigns higher weights to terms that are frequent in a specific document but rare across the entire corpus, indicating their significance in distinguishing that document's content.

The resulting TF-IDF values form a two-dimensional matrix $V \in \mathbb{R}^{D\times T}$, where D is the number of documents and T is the number of unique terms (features) in the corpus. This matrix serves as the input for machine learning algorithms, with each document represented as a vector of its TF-IDF features. In this study, both unigrams and bigrams are used to construct the feature space, providing richer contextual information for subsequent analysis.

3.4.4 Topic Modeling with NMF

Topic modeling was conducted using Non-Negative Matrix Factorization (NMF), an unsupervised dimensionality reduction technique that is particularly effective for uncovering latent thematic structures within large textual datasets. NMF decomposes the input TF-IDF matrix $V \in \mathbb{R}^{D\times T}$, where D is the number of documents and T is the number of unique terms, into two non-negative matrices:

- **Document-topic matrix** ($W \in \mathbb{R}^{D \times k}$): Represents the prevalence or distribution of each topic across individual documents.

- **Topic-term matrix** ($H \in \mathbb{R}^{k \times T}$): Indicates the relevance or contribution of specific terms to each identified topic.

Formally, NMF seeks to approximate the original matrix V through the factorization $V \approx WH$, minimizing the reconstruction error as measured by the Frobenius norm:

$$\min_{W,H \geq 0} \|V - WH\|_F^2$$

where k denotes the number of latent topics.

Selecting an appropriate number of topics (k) is crucial for ensuring the interpretability and practical relevance of the model. In this study, the optimal value of k was determined using the C_v coherence metric, as proposed by Röder, Both, and Hinneburg (2015) [16]. The C_v metric quantitatively assesses the semantic consistency of terms within topics, based on their co-occurrence patterns across the corpus. While coherence scores provided valuable quantitative guidance, the final selection of k also incorporated human judgment to ensure interpretability and to reduce redundancy among topics.

3.4.5 Topic-Specific Sentiment Analysis

To conduct detailed, topic-specific sentiment analysis, a zero-shot classification approach was employed using the **joeddav/xlm-roberta-large-xnli** transformer model available on the Hugging Face platform [17]. This model is built upon the XLM-RoBERTa-large architecture a robust, multilingual extension of RoBERTa, pre-trained on a large-scale corpus covering over 100 languages. The XNLI (Cross-lingual Natural Language Inference) fine-tuning further equips the model to perform semantic understanding and natural language inference across languages, rendering it highly effective for multilingual and cross-cultural text analysis.

The model's zero-shot classification capability stems from its fine-tuning on the XNLI dataset, enabling it to categorize text into user-defined topics by reframing classification as an entailment problem [17]. Specifically, the model determines whether a given text segment "entails" a hypothesis statement representing a target thematic label (e.g., "This text is about transportation").

The methodological workflow was as follows:

1. Each review was segmented into individual sentences.

2. Each sentence was classified into one or more predefined thematic categories using the zero-shot classification approach.

3. The positive sentiment score for each sentence was computed using the sentiment analysis model (**cardiffnlp/twitter-xlm-roberta-base-sentiment**).

4. The results were aggregated to yield:

   - The overall positive sentiment score for the full review,
   - A mapping of sentences labeled by thematic category,
   - Topic-specific sentiment scores for in-depth thematic analysis.

Leveraging transformer-based models pre-trained on multilingual, real-world data enabled the analytical workflow to bypass additional text pre-processing steps for sentiment analysis (e.g., punctuation removal, stop-word filtering), thereby streamlining the process and preserving semantic integrity.

3.5 Assessing Potential Local Bias (H2)

To assess whether there is any bias in review sentiment between Kazakhstanis and international tourists, the following statistical tests were applied.

3.5.1 Mann-Whitney U-test

To quantitatively compare sentiment distributions between reviews authored by Kazakhstanis and those written by international tourists, the Mann-Whitney U-test was employed. Unlike parametric tests such as the t-test, the Mann-Whitney U-test does not assume normality of the data, making it particularly well-suited for sentiment scores, which is shown to deviate from normal distribution (see Figure 5.3).

The Mann-Whitney U-test evaluates whether there is a significant difference between two independent groups by comparing the ranks of their observations. The test statistic U is calculated as follows:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

where R_1 and R_2 are the sums of ranks for group 1 (Kazakhstanis) and group 2 (international tourists), respectively, and n_1 and n_2 are the corresponding sample sizes. The smaller of U_1 and U_2 is used for computing the p-value.

The test assesses the following hypotheses:

- Null hypothesis (H_0): The distributions of sentiment scores for the two groups are the same.

- Alternative hypothesis (H_1): The distributions of sentiment scores for the two groups are different.

A low p-value indicates a significant difference in sentiment distributions between the two groups, which may suggest potential bias if Kazakhstanis systematically display higher sentiment scores.

The Mann-Whitney U-test requires the following four assumptions:

1. The dependent variable (sentiment score) is continuous or ordinal.
2. The independent variable is categorical with two clearly defined groups (Kazakhstanis vs. international tourists).
3. Observations within and between groups are independent.
4. Samples are drawn representatively from their respective populations.

3.5.2 Effect Size: Cohen's *d*

While statistical significance indicates whether a difference exists between groups, it does not reflect the practical importance of this difference. Therefore, Cohen's d was computed as a measure of effect size. Cohen's d quantifies the standardized difference between the means of two groups and is calculated as follows:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups, and s_p is the pooled standard deviation, defined as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Here, s_1 and s_2 are the standard deviations and n_1 and n_2 are the sample sizes of the two groups.

Cohen's conventional thresholds 0.2 (small effect), 0.5 (medium effect), and 0.8 (large effect) were used to interpret the magnitude of the observed effect size.

3.6 Assessing Potential Seasonality of Sentiment (H3)

To investigate whether tourist sentiment exhibits significant seasonal patterns, the following time series and statistical analyses were conducted.

3.6.1 Time series decomposition and the Kruskal-Wallis test

Examining the seasonality of positive sentiment scores is essential, as it reveals recurring patterns in tourist feedback throughout the year. Understanding these cyclical trends allows for more accurate forecasting and supports targeted planning for interventions or marketing strategies during periods of consistently higher or lower sentiment.

To rigorously analyze temporal patterns, this study employs both **time series decomposition** and the **Kruskal-Wallis test**. Time series decomposition separates the observed sentiment time series into three primary components: trend, seasonality, and residual (noise). The *trend* component captures long-term progression, the *seasonality* component isolates periodic fluctuations (monthly patterns), and the *residual* component represents random variation not explained by trend or seasonality. The additive decomposition model is expressed as:

$$y_t = T_t + S_t + R_t$$

where

- y_t is the observed sentiment score at time t,
- T_t is the trend component,
- S_t is the seasonal component, and
- R_t is the residual (random noise) component.

In addition, the Kruskal-Wallis test a non-parametric method for comparing the distributions of three or more independent groups is applied to statistically assess whether significant differences in positive sentiment scores exist across different months. Unlike traditional ANOVA, the Kruskal-Wallis test does not assume normality and is robust to outliers and heteroscedasticity, making it well suited for sentiment data. A significant result from this test indicates the presence of meaningful seasonality, confirming that sentiment scores systematically vary between months.

The Kruskal-Wallis test statistic is calculated as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i^2\right) - 3(N+1)$$

where

- N is the total number of observations,
- k is the number of groups (months),
- n_i is the number of observations in group i,

- $\bar{R}_i$ is the sum of the ranks in group i.

Together, these methods provide a comprehensive framework for detecting and quantifying seasonal effects in sentiment data.

Chapter 4

Data

This section describes the main features of the dataset and provides an initial exploration of its structure and key characteristics.

4.1 Descriptive Statistics and Exploratory Analysis

The final dataset contains 23,296 tourist reviews and 18 selected features. The reviews span from 2011 to early 2025, though reviews from 2025 were not included in the analysis because data collection took place early in the year, which could have resulted in an incomplete and unrepresentative sample for 2025.

An analysis of review frequency over time reveals the following pattern: from 2011 to 2016, there is a steady increase in the number of reviews. This trend reverses between 2016 and 2019, with a noticeable decline. A sharp drop in 2020 follows, likely reflecting the global travel restrictions and reduced tourism caused by the COVID-19 pandemic. From 2021 to 2024, there is a gradual recovery, with a slow upward trend in review counts (see Figure 4.1). This pattern aligns closely with the official tourism statistics from the Bureau of National Statistics of Kazakhstan, which reports the annual number of incoming tourists. While both datasets follow a similar overall trend, the official figures exhibit greater volatility and display a negative correlation with TripAdvisor review counts during 2016–2018 (see Figure 5.1). One plausible explanation for this divergence is a potential decline in TripAdvisor's popularity as a review platform during that period. Further investigation is warranted, particularly focusing on incoming tourism only.

The variable tripType provides insight into the context of travel. Among the available categories, the most common trip type is ”with friends”, while ”business trips” are the least frequently mentioned (see Figure 4.2). This is consistent with the assumption that tourists using TripAdvisor are more likely to be leisure travelers than business visitors.

The dataset comprises reviews in 26 different languages. However, Russian and English dominate, together accounting for approximately 20,000 out of the 23,296 reviews (see Fig-

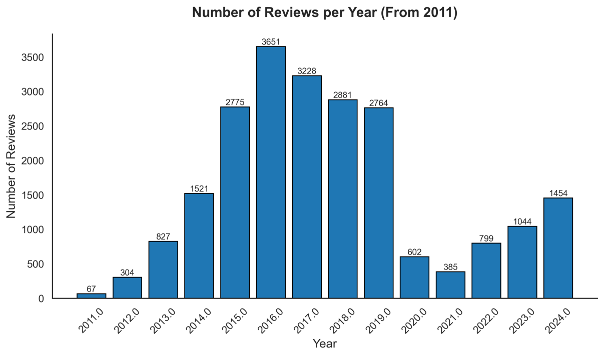

Figure 4.1: Number of reviews per year

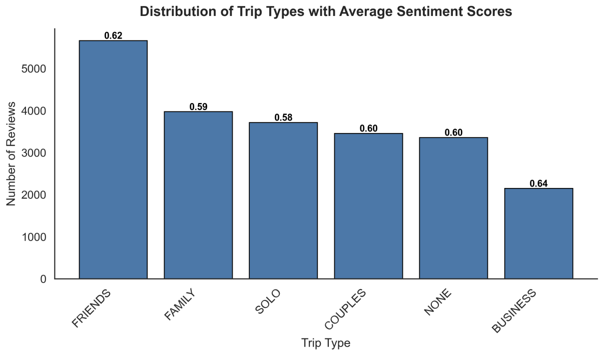

Figure 4.2: Distribution of trip types among tourists

ure 4.3). This reflects the regional linguistic landscape and the global usage of English in tourism-related content.

Analysis of user metadata reveals that the top countries of origin for reviewers visiting Kazakhstan include Russia, United Kingdom, Italy, India, United Arab Emirates, Germany, Japan, Australia, The Netherlands (see Figure 4.4). This geographic spread indicates that Kazakhstan attracts a diverse set of international tourists, with particularly strong representation from Russia and European countries.

Analysis of the time difference between the travel date and the published date of TripAdvisor reviews reveals a pronounced recency effect in reviewer behavior. The vast majority of reviews are submitted within the first month following a visit, and the number of reviews declines sharply as the interval between travel and review publication increases. Very few trav-

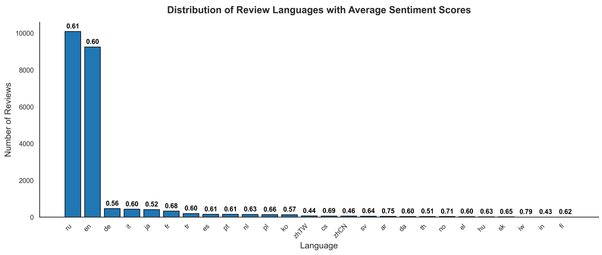

Figure 4.3: Distribution of languages

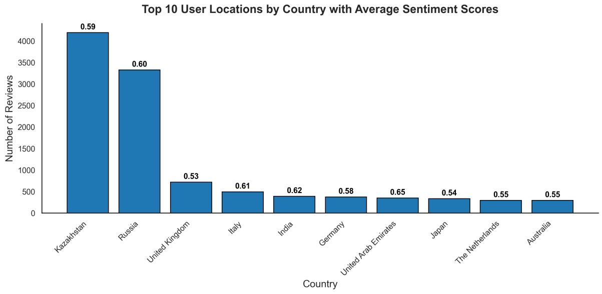

Figure 4.4: Top 10 countries of tourists

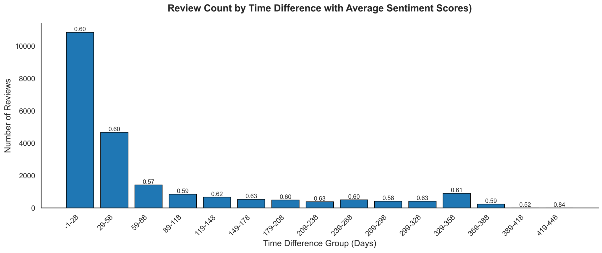

Figure 4.5: Distribution of time difference between travel date and published date

elers choose to write reviews more than a few months after their experience, with review counts dropping to almost negligible levels beyond six months (see Figure 4.5). This distribution indicates that tourists are most inclined to share their impressions soon after their trip, likely while their memories and perceptions are still fresh. Reviews written a significant time after the travel date are rare, and their limited frequency suggests that they are not representative of the broader reviewing population. Overall, the data demonstrate that user-generated content on TripAdvisor primarily reflects recent visitor experiences, which is important to consider when interpreting trends and patterns in online review data.

The number of reviews per attraction follows a power-law distribution (see Figure 4.6). A small subset of attractions receives a disproportionately large number of reviews, while the vast majority are reviewed infrequently. This skew is expected and reflects the phenomenon where a few iconic or heavily promoted attractions (e.g., major landmarks or capital city sites) serve as focal points for tourism, thereby shaping international perceptions of the country [18] [19].

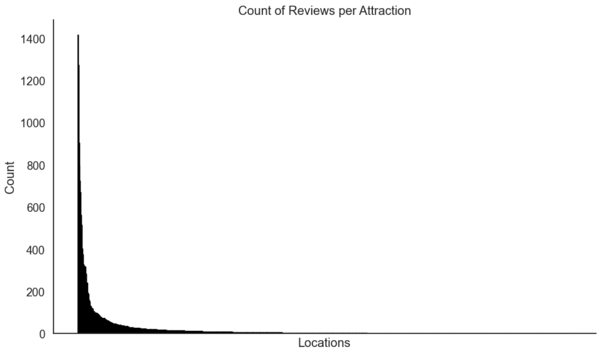

Figure 4.6: Distribution of the number of attraction reviews

Chapter 5

Results

This chapter presents the main findings of the study, organized around the central research question and hypotheses. It starts by exploring the relationship between TripAdvisor review activity and official records of tourist arrivals. The results then progress through analyses of sentiment trends, topic modeling outcomes, and statistical tests of local bias and sentiment seasonality.

5.1 Correlation Between TripAdvisor Reviews and Tourist Arrivals

A Pearson correlation analysis demonstrated a statistically significant positive association between the number of TripAdvisor reviews and official international tourist arrivals in Kazakhstan ($r = 0.663$, $p = 0.0136$) (see Figure 5.1). This finding indicates that user-generated review activity on TripAdvisor may serve as a moderately strong proxy for real-world tourism trends in Kazakhstan. In other words, increased engagement on TripAdvisor appears to be reliably linked with higher numbers of tourists visiting the country, underscoring the potential of digital trace data as a supplementary indicator for tourism analysis.

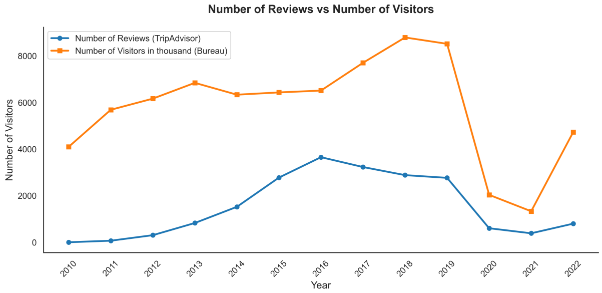

Figure 5.1: Number of reviews from TripAdvisor and number of visitors from Bureau

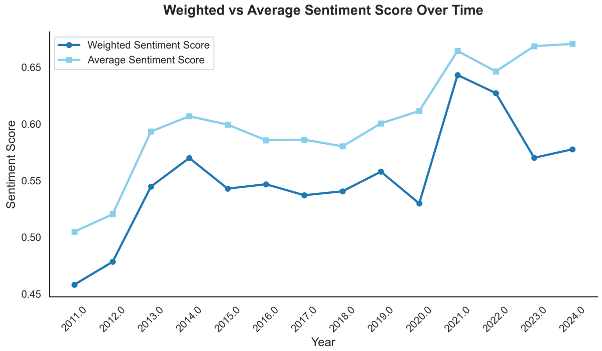

Figure 5.2: Weighted and average sentiment score over time

5.2 Sentiment Analysis Results (H1)

This section presents the results of the sentiment analysis, including overall trends from 2011 to 2024, weighted sentiment scores, the distribution of sentiment values, and differences across regions of Kazakhstan.

5.2.1 Overall Sentiment Trends (2011–2024)

Sentiment analysis of tourist reviews for Kazakhstan's attractions indicates a clear and generally positive trajectory in average sentiment scores from 2011 to 2024 (see Figure 5.2). This upward trend suggests a progressive improvement in visitor satisfaction over the period, which may reflect the ongoing development and modernization efforts within Kazakhstan's tourism sector.

Although sentiment scores fluctuated during the earlier years of the period under study, the interval from 2014 to 2020 is marked by relative stability. Notably, a pronounced increase in average positive sentiment is observed after 2020, pointing to a substantial shift toward more favorable tourist experiences. This recent improvement is likely attributable to increased governmental investment in tourism infrastructure, including significant enhancements to transportation networks and improved accessibility to natural sites. These trends align with the launch and ongoing implementation of Kazakhstan's State Program for the Development of the Tourism Industry (2019–2025), which prioritized infrastructure upgrades, road construction, airport modernization, and initiatives aimed at improving access to key tourist destinations [20] [21]. Enhanced infrastructure and greater accessibility have been highlighted in policy documents and independent reports as crucial factors in improving the quality of tourist experiences and shaping international perceptions of Kazakhstan as a travel destination.

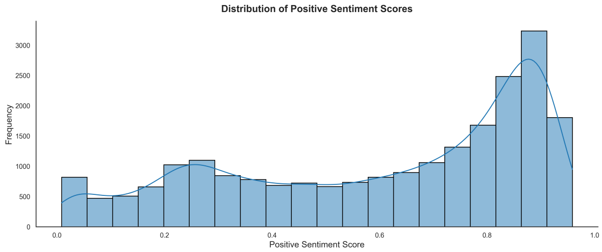

Figure 5.3: Distribution of positive sentiment score

5.2.2 Weighted Sentiment Scores

To better reflect the perceived credibility and influence of individual reviews, a weighted average positive sentiment score was calculated. The weighting scheme assigns greater importance to reviews that have received more helpful votes from other users: each review is assigned a base weight of 1, with each additional helpful vote incrementing the weight by 1 (e.g., a review with three helpful votes receives a weight of 4). This approach prioritizes reviews that readers have deemed informative or trustworthy.

A strong positive correlation was observed between the unweighted and weighted sentiment averages over time, suggesting overall consistency between the two measures (see Figure 5.2). However, the weighted sentiment scores display greater volatility, particularly between 2019 and 2024. Furthermore, the weighted average consistently trails the unweighted average, indicating that reviews with lower sentiment scores tend to receive more helpful votes. This finding implies that tourists are more inclined to endorse or value critical feedback, perhaps because such reviews are perceived as more balanced or informative.

5.2.3 Distribution of the Sentiment Score

The distribution of positive sentiment scores, as depicted in the figure above (see Figure 5.3), illustrates a nuanced pattern. Sentiment scores are relatively evenly distributed in the range from 0 to 0.75. However, there is a pronounced peak in the 0.75 to 1.00 interval, indicating that a substantial proportion of reviews exhibit high levels of positive sentiment. This suggests that while tourist experiences in Kazakhstan vary, there is a strong tendency toward highly favorable evaluations.

Upon further inspection, the distribution exhibits two notable peaks: one around 0.25 and another around 0.90. This bimodal pattern indicates that, while a considerable number of reviews report lower levels of positive sentiment, a significantly larger share clusters at the upper end of the scale. This skew towards higher sentiment scores is particularly prominent in the more recent years analyzed.

A year-by-year examination of sentiment distributions shows that, although the overall pattern remains relatively stable across most years, there has been a clear increase in the prevalence of high positive sentiment scores, especially since 2019–2020 (see Figure 5.4). This period is

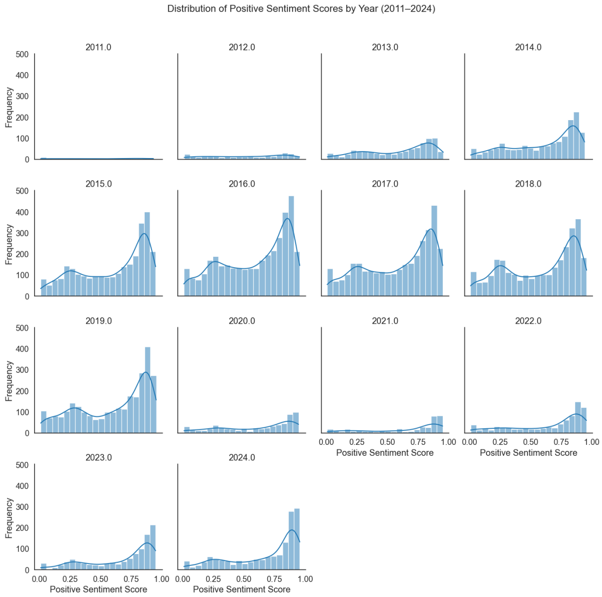

Figure 5.4: Distribution of positive sentiment score by year

characterized by both a rise in average sentiment and a reduction in the dispersion of sentiment scores, suggesting that negative or mixed reviews have become less common in recent years.

To further illustrate the overall trends in tourist sentiment, selected review excerpts provide additional context to the quantitative findings. Examples of low-sentiment reviews often highlight recurring frustrations with value, service quality, or infrastructure. For instance, one visitor expressed disappointment with the region's ski facilities: "They refurbished the area for Asian Winter Olympics but c'mon guys this is still not a worldwide ski center... Tickets, skipasses, rentals, everything is too expensive. I'd rather go Alpines instead." Another review pointed to inconsistent service and cleanliness: "Advertising in the cabs doesn't allow to see beautiful scene. Cabs not cleaned well." Complaints about visitor management are also evident, such as a museum visitor who noted: "DO NOT bring the kids with you. Especially babies. Because everybody will be yelling at you for bringing kids and then follow you from room to room and yelling more. I felt like a repression victim myself." In addition, issues with restrictive policies and unfriendly staff are mentioned: "The most beautiful church in Almaty is actually that of Zenkov. Unlike the building, the staff is less friendly, we were not allowed to take pictures inside which is a shame to make this place known abroad."

In contrast, examples of high-sentiment reviews frequently emphasize natural beauty, positive experiences with guides, and memorable atmospheres. One visitor shared: "Snow capped mountains were mesmerizing and our guide was really helpful and made our trip really memorable." Another praised the modernity and quality of a museum: "Everything was really impressive and it is definitely worth a visit!" Several reviews highlighted the peacefulness and attractiveness of urban parks and public spaces: "It's a wonderful place to go back again and again, with a breath of peace and tranquility, a beautifully equipped fountain of holy water, a stunning temple." Positive feedback also references convenience and value, as seen in: "The

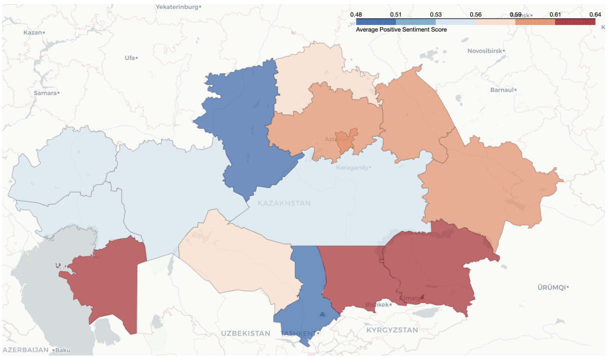

Figure 5.5: Average positive sentiment score by regions

cottage elevator is very convenient, it's cheaper, but it's also more interesting."

5.2.4 Regional Sentiment Distribution

A spatial analysis was performed by aggregating average positive sentiment scores across Kazakhstan's 16 administrative regions (see Figure 5.5). The results reveal that the southern region has many of the country's flagship tourist attractions, including prominent ski resorts, mountain ranges, and alpine lakes, consistently report the highest average sentiment scores. This aligns with expectations, as these areas attract the majority of both domestic and international tourists.

Conversely, west regions exhibit significantly lower volumes of reviews, limiting the statistical robustness of sentiment estimates for these areas. The scarcity of data points from west Kazakhstan precludes definitive conclusions regarding tourist satisfaction in these regions. The observed spatial disparities highlight the importance of regional tourism development and the need for targeted strategies to boost both visitation and tourist satisfaction outside the primary tourist corridors.

5.3 Topic Modeling Results (H1)

As described in Section 3.4.4, Non-negative Matrix Factorization (NMF) was applied to the corpus of review texts to identify prominent topics discussed by tourists. The optimal number of topics (k) was initially determined using the coherence score (C_v) as an objective metric, with supplementary human interpretation to ensure meaningful thematic separation [22]. The evolution of coherence scores across different k values is presented in Figure 5.6.

Although k = 13 yielded the highest coherence score ($C_v = 0.645$), it resulted in significant thematic overlap between topics. After further experimentation and consideration of interpretability, a final model with k = 7 topics was selected, achieving a comparable coherence score ($C_v = 0.635$) while ensuring clearer separation of themes. These seven topics provide a concise yet comprehensive overview of the main aspects highlighted in tourist reviews, ranging from natural attractions and service quality to infrastructure and accessibility [22].

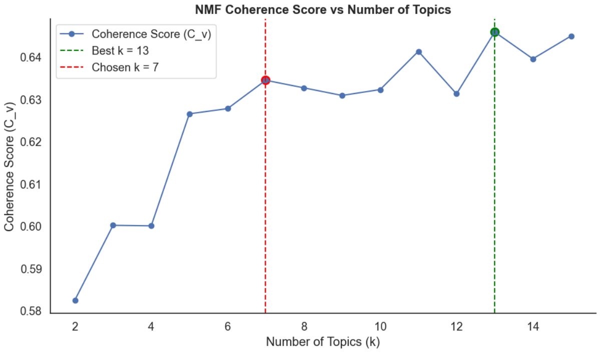

Figure 5.6: NMF coherence score VS number of topics

The extraction of the top 20 n-grams from the topic-term matrix enabled the identification of seven dominant themes within the corpus of tourist reviews (see Table 5.1). These themes encapsulate the main areas of focus and interest expressed by visitors to Kazakhstan. The Transportation topic includes terms such as “car,” “cable car,” “bus,” “mountain,” and “ride,” indicating frequent discussion of travel logistics, local transport options, and the experience of reaching key attractions often in and around Almaty. Tourism Services is characterized by words like “tour,” “guide,” “recommend,” “friendly,” and “driver,” reflecting a strong emphasis on guided experiences, the quality of tour operations, and interactions with service personnel. The Culture and History theme is signaled by n-grams such as “museum,” “exhibit,” “Kazakh,” “Soviet,” and “art,” suggesting that many reviews center on visits to museums, historical buildings, and cultural events, as well as the representation of national identity and heritage.

Nature and Hiking emerges through terms like “lake,” “mountain,” “canyon,” “nature,” “walk,” and “hike,” highlighting Kazakhstan’s appeal as a destination for natural scenery, outdoor activities, and adventure tourism. The Religious and Spiritual Sites topic features n-grams including “mosque,” “cathedral,” “architecture,” “temple,” and “prayer,” capturing reviews related to visits to religious landmarks, architectural appreciation, and spiritual experiences. Urban Parks is identified through words such as “park,” “monument,” “fountain,” “memorial,” “tree,” and “attraction,” indicating interest in green spaces, recreational areas, monuments, and city landmarks, particularly in Almaty and Astana. Lastly, the Food and Leisure topic is characterized by terms such as “food,” “mall,” “relax,” “shop,” “price,” and “restaurant,” pointing to reviews discussing dining, shopping, relaxation, family activities, and value for money.

Together, these themes provide a comprehensive overview of the diverse aspects of tourist experiences in Kazakhstan, from practical concerns about transportation and services to cultural, natural, and recreational attractions, and everyday leisure activities. The prevalence of these topics underscores the multifaceted nature of tourism in Kazakhstan as reflected in visitor feedback.

If we look at the topic modeling deeper into the years, we can highlight three topics that dominate in each year from 2011 to 2024 (see Table 5.2).

Analysis of the prevalence of these topics over time showed that Food and Leisure was a

No.	Topic	Top Terms
1	Transportation	view, city, car, cable, cable car, mountain, Almaty, ride, hill, tower, great view, restaurant, zoo, beautiful view, great, bus, climb, tenge, café, kok, ski, enjoy, Beatles
2	Tourism Services	tour, guide, trip, day, recommend, Almaty, great tour, experience, canyon, amazing, Kazakhstan, friendly, highly recommend, best, driver, knowledgeable, excellent, thanks, Charyn, English, wonderful, company, enjoyed
3	Culture & History	museum, history, interesting, Kazakhstan, exhibit, building, exhibition, Kazakh, hall, instrument, visit, Astana, modern, art, collection, English, worth, national, Soviet, culture, people, Russian, entrance, president
4	Nature & Hiking	lake, mountain, road, water, Almaty, canyon, beautiful, drive, Almaty Lake, hour, Kolsai, nature, km, walk, Big Almaty Lake, river, car, Charyn, beauty, hike, snow, Kaindy, trip, air
5	Religious & Spiritual Sites	beautiful, mosque, inside, building, church, cathedral, Astana, architecture, beautiful place, worth, look, interior, dome, temple, Orthodox, impressive, city center, largest, prayer, central
6	Urban Parks	park, monument, cathedral, walk, located, Panfilov, war, memorial, church, tree, city, Almaty, center, fountain, green, area, attraction, statue, hero, Orthodox, Soviet, child, large, walking, amusement
7	Food & Leisure	place, good, nice, great, time, lot, visit, kid, walk, food, family, shopping, people, mall, relax, shop, price, clean, spend, winter, restaurant, child

Table 5.1: Topic modeling results: Identified topics and their top terms

Year	Dominant Topics
2011	Food & Leisure, Transportation, Nature & Hiking
2012	Food & Leisure, Urban Parks, Nature & Hiking
2013	Food & Leisure, Transportation, Culture & History
2014	Food & Leisure, Religious & Spiritual Sites, Culture & History
2015	Food & Leisure, Religious & Spiritual Sites, Culture & History
2016	Food & Leisure, Culture & History, Religious & Spiritual Sites
2017	Food & Leisure, Culture & History, Religious & Spiritual Sites
2018	Food & Leisure, Culture & History, Tourism Services
2019	Tourism Services, Food & Leisure, Nature & Hiking
2020	Tourism Services, Food & Leisure, Nature & Hiking
2021	Tourism Services, Food & Leisure, Nature & Hiking
2022	Tourism Services, Food & Leisure, Nature & Hiking
2023	Tourism Services, Nature & Hiking, Food & Leisure
2024	Tourism Services, Nature & Hiking, Food & Leisure

Table 5.2: Top three dominant topic modeling results by year.

dominant topic in every year from 2011 to 2024. From 2011 to 2017, Culture and History, and Religious and Spiritual Sites also appeared frequently as dominant topics. The transportation topic was prevalent in 2011 and 2013. Beginning in 2018, there was a shift, with Tourism Services and Nature and Hiking increasingly becoming dominant topics alongside Food and Leisure. From 2019 through 2024, Tourism Services and Nature and Hiking appeared most frequently among the top three topics each year.

Topic	**Reviews Mentioning (% of Total)**
Food & Leisure	35.8% (8,331)
Tourism Services	71.6% (16,679)
Culture & History	69.9% (16,273)
Nature & Hiking	59.6% (13,893)
Transportation	52.6% (12,248)
Urban Parks	38.9% (9,063)
Religious & Spiritual Sites	52.0% (12,125)

Table 5.3: Percentage of topics mentioned in tourist reviews.

According to the analysis, the tourism services topic is the most prevalent, appearing in 71.6% of all reviews (see Table 5.3). In contrast, the food and leisure topic shows the lowest overall coverage, with 35.8% of reviews, although it remains one of the dominant topics each year. This result is likely due to a more even distribution of food and leisure mentions across the dataset compared to other topics. Urban parks have a coverage rate of 38.9%, with most mentions concentrated in the period 2011–2014, when the total number of reviews was relatively low. All other identified topics have coverage rates exceeding 50%.

Following the assignment of topic labels to individual sentences, a positive sentiment score was calculated for each sentence. Analysis of sentiment scores across topics for the period 2013–2024 shows that most topics including food and leisure, culture and history, religious and spiritual sites, tourism services, and urban parks have average positive sentiment scores below the overall mean. The transportation topic consistently exhibits an average positive sentiment score that is considerably lower than both the overall average and the averages of other topics. This indicates that reviews mentioning transportation contribute disproportionately to lowering the total average positive sentiment score across all years and are associated with a higher volume of negative sentiment (see Figure 5.7).

In contrast, the nature and hiking topic generally maintains average positive sentiment scores above the overall mean for most years in the study period. Reviews in this category frequently mention positive aspects such as beautiful views, mountains, and natural attractions including canyons and lakes.

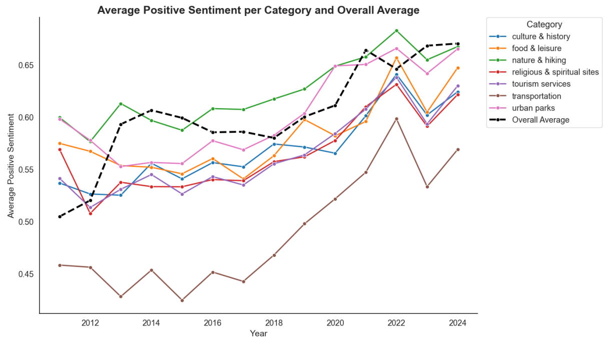

Figure 5.7: Average positive sentiment score per topic by year

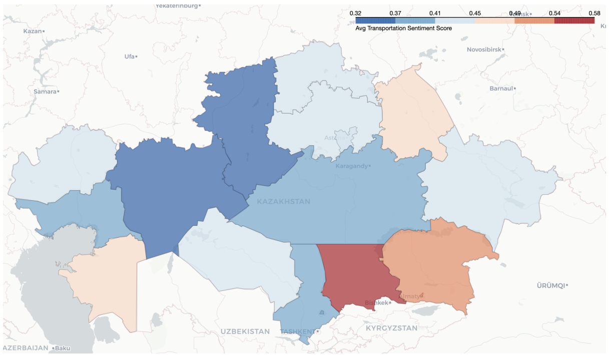

Figure 5.8: Average positive sentiment score of transportation related sentences by regions

5.3.1 Sentiment by Transportation Topic

Analysis using zero-shot topic classification and transformer-based segmentation shows that transportation is referenced in 52.6% of all tourist reviews. The average positive sentiment score for transportation-related segments is consistently and significantly lower than the overall sentiment score for the entire review dataset.

Regional analysis of sentiment reveals that transportation-related reviews from the southern region of Kazakhstan home to the largest cities and key tourist destinations exhibit the highest average positive sentiment scores. In contrast, most other regions display relatively low average positive sentiment scores in transportation-related reviews (see Figure 5.8).

Examination of sample transportation-related review sentences highlights the centrality of cable cars, gondola rides, and associated infrastructure in tourists' experiences and feedback. Tourists frequently comment on the multiple stages of cable car rides required to reach mountain summits, ticketing complexities (e.g., "gondola 360 + combi 1 + combi 2" packages), the duration of these journeys, and the overall efficiency and scenic value of the cable car system. While some reviews note smooth operation and pleasant scenery, others mention long waiting times and underutilized slopes. In addition to these infrastructure-focused comments, reviews often enumerate the range of amenities available at transport hubs, such as elevators, ski schools, repair points, restaurants, shops, and parking facilities. Few examples of transportation-related reviews:

- “There are 3 cable car rides to ride all the way to the top of the mountain.”
- “You need a ticket that is under the ‘gondola 360 + combi 1 + combi 2’ package.”
- “The rides to the 3 different stations are quite long, especially the gondola ride.”
- “The cable car system ran smoothly and provided good scenery along the way.”
- “We had to wait long for our cable car.”
- “A few skiers held us company down below, but the slopes were mostly empty.”
- “It has cable car, elevators, ski school, first aid station, equipment rental and repair point, restaurant and cafe, shops, parking, etc.”

Top transportation-related words	**Interpretation**
car, bus, road	These core transport words indicate that issues may center around vehicle quality, availability, traffic, or road infrastructure.
Almaty, lake, city, place	Location-specific mentions suggest negative transportation experiences were associated with travel to or within Almaty, lake regions, or urban areas possibly due to distance, road conditions, or navigation issues.
walk, went, way, going	These verbs imply difficulty reaching destinations, or being forced to walk due to poor transportation options.
just, don, didn, people	These function words often appear in frustrated or informal phrasing (e.g., “people didn’t show up”, “don’t take the bus”, “just terrible”), revealing emotional tone even in n-gram form.
time, hours, day	Mentions of time may signal delays, long travel durations, or poor scheduling.
cable	Possibly refers to cable cars or ski lifts could indicate poor experiences in mountain or nature tourism spots.
tour	Suggests dissatisfaction with organized tours, possibly due to transport logistics or mismanagement.

Table 5.4: Topic modeling results: Interpretation of top transportation-related words in tourist reviews

Sentences with lower sentiment scores commonly reference poor road quality, lengthy travel times, issues with vehicle types, and limited transportation options. Location-specific complaints are frequent, particularly in relation to Almaty and lake regions, where travelers report difficulties reaching destinations, reliance on walking due to inadequate transport, and other logistical frustrations. Analysis of the most common transportation-related words further supports these findings: terms such as “car,” “bus,” and “road” point to concerns over vehicle quality, traffic, and infrastructure; mentions of “Almaty,” “lake,” and “city” identify

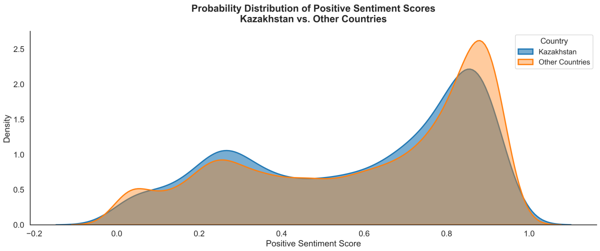

Figure 5.9: Probability distribution of positive sentiment scores of Kazakhstanis and incoming tourists

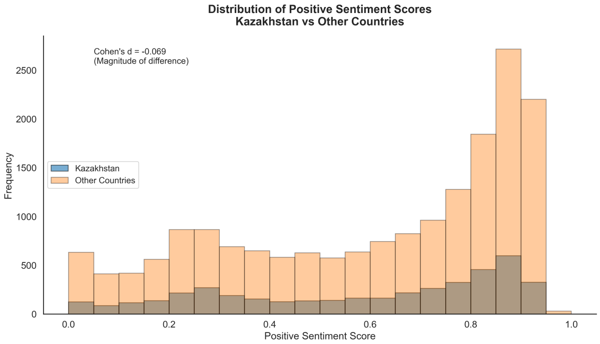

Figure 5.10: Distribution of positive sentiment scores of Kazakhstanis and incoming tourists

problematic travel corridors; and verbs like “walk,” “went,” and “going” indicate challenges in accessing sites. The frequent use of function words such as “just,” “don,” and “didn’t” reflects an informal, sometimes frustrated tone in the reviews, while references to “time,” “hours,” and “day” often signal dissatisfaction with delays or scheduling. Finally, the prominence of the word “tour” in low-sentiment segments suggests that organized tours are not immune to transport-related critiques, with complaints sometimes linked to logistics or mismanagement (see Table 5.4). Collectively, these findings demonstrate that transportation issues, both infrastructural and operational, are a significant source of negative sentiment in visitor feedback.

5.4 Kazakhstanis Bias Analysis (H2)

A Mann-Whitney U test was performed to evaluate whether the distribution of positive sentiment scores differs between reviews authored by Kazakhstanis and those written by reviewers from other countries (see Figure 5.9). The analysis yielded a statistically significant result ($U = 35,838,918$, $p < 0.001$). The effect size, measured by Cohen’s d, was -0.069, indicating a negligible difference in positive sentiment scores between the two groups (see Figure 5.10). These results demonstrate that, while a statistical difference is present, the practical magnitude of this difference is minimal. Thus, no meaningful bias is observed in the sentiment scores of reviews authored by Kazakhstanis compared to those written by international tourists.

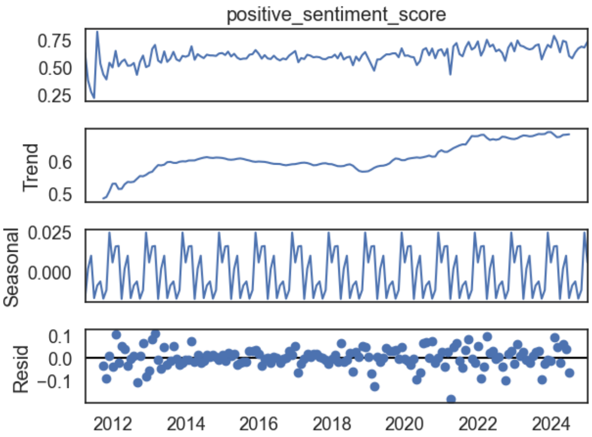

Figure 5.11: Time series decomposition of positive sentiment scores

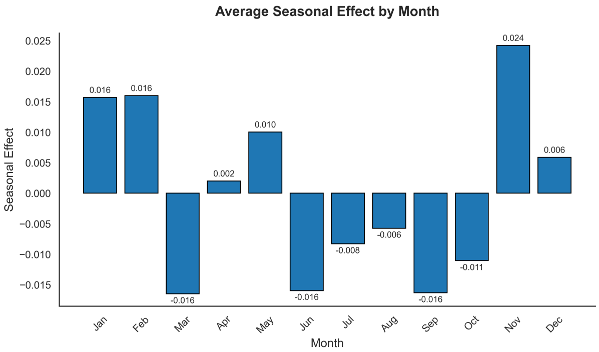

Figure 5.12: Average seasonal effect by month

5.5 Seasonality Check Results (H3)

Time series decomposition of positive sentiment scores revealed three main components: trend, seasonality, and residual variation. The trend component indicated a steady upward movement in sentiment scores from 2012 to 2024, with observable periods of acceleration and stabilization. The seasonal component identified a consistent annual cycle, with sentiment scores rising and falling at similar points each year, demonstrating a recurring monthly seasonality. The residual component showed small, randomly distributed variations around zero, suggesting that most variability is accounted for by the trend and seasonal components (see Figure 5.11).

Analysis of the seasonal component showed regular month-to-month variation. Sentiment scores peaked in November and reached their lowest point in March each year. To statistically validate the presence of seasonality, the Kruskal-Wallis test was applied to sentiment scores by month. The test result was highly significant ($H - statistic = 60.739$, $p < 0.0001$), confirming that the differences in sentiment scores across months are statistically significant and not due to random variation. These findings demonstrate that positive sentiment scores are affected by both long-term trends and recurring seasonal effects, with the highest scores observed in November and the lowest in March (see Figure 5.12).

Chapter 6

Discussion

This chapter interprets and contextualizes the key results of the analysis, connecting them to the research questions and the broader literature on tourism analytics. The discussion explores how sentiment trends, topic modeling outcomes, and statistical findings on bias and seasonality contribute to our understanding of visitor experiences and tourism development in Kazakhstan.

6.1 Digital Reviews Data as a Proxy for Tourism Flows

The strong positive correlation observed between the number of TripAdvisor reviews and official tourist arrival statistics underscores the potential of digital trace data as an effective proxy for monitoring real-world tourism activity in Kazakhstan. This finding is consistent with a growing body of research demonstrating that online platforms, and particularly user-generated content, can serve as valuable supplements to traditional data sources for tracking trends in tourism demand [23]. In contexts where timely or granular official statistics are not always available, digital traces provide a near real-time lens on visitor behavior, preferences, and sentiment.

In the case of Kazakhstan, the moderate-to-strong correlation (r = 0.663) suggests that spikes in TripAdvisor review activity align closely with increases in physical tourist arrivals. This result reinforces the validity of using large-scale review data for tourism research and policymaking, especially in emerging tourism markets. However, while digital trace data reflects aggregate visitor flows, it may not capture all visitor segments equally, particularly those less likely to engage with online platforms. As such, these data should be interpreted as complementary rather than fully substitutive to official statistics.

6.2 Sentiment Trends and Their Implications

The longitudinal analysis of sentiment scores from TripAdvisor reviews indicates a steady improvement in tourist sentiment toward Kazakhstan's attractions from 2011 through 2024. This upward trend is particularly pronounced after 2020, where both the mean positive sentiment and the consistency of sentiment scores increased. These findings suggest ongoing enhancements in the tourism experience, likely resulting from sustained investments in infrastructure, hospitality, and service quality.

The identification of stable and, in recent years, increasingly positive sentiment aligns with reports of national efforts to modernize the tourism sector. Targeted improvements in transportation, accommodation, and the accessibility of natural attractions may have contributed to the observed rise in satisfaction among visitors. The decline in sentiment dispersion in later years further indicates that the gap between highly positive and negative experiences has narrowed, suggesting greater uniformity in the quality of tourism offerings.

These developments correspond with the introduction and continued implementation of Kazakhstan's State Program for the Development of the Tourism Industry (2019–2025), which emphasized modernizing infrastructure, constructing new roads, upgrading airports, and improving access to prominent tourist sites [20] [21]. Policy documents and independent analyses have consistently identified improved infrastructure and increased accessibility as key drivers behind higher visitor satisfaction and enhanced international perceptions of Kazakhstan as a travel destination.

6.3 Sentiment by Topic

The topic-based analysis reveals significant variation in sentiment across different aspects of the tourism experience. Among all topics, transportation stands out as the area most consistently associated with lower positive sentiment. With transportation mentioned in over half of all reviews (52.6%), and with average positive sentiment scores for transportation-related segments significantly below the overall mean, the results highlight transportation as a persistent challenge for Kazakhstan's tourism industry. This pattern is visible across most regions, with particularly low sentiment in areas outside the well-developed south. Frequent complaints include references to road conditions, delays, the quality and availability of vehicles, and challenges related to travel logistics.

By contrast, reviews related to nature and hiking display the highest positive sentiment scores. These segments consistently exceed the overall average, suggesting that natural attractions such as mountains, canyons, and lakes generate especially favorable visitor experiences. This finding underscores the strength of Kazakhstan's natural assets as a core component of its tourism appeal.

The regional analysis further reinforces these findings. The southern region, home to major cities and flagship attractions, records both the highest volumes of reviews and the most positive transportation sentiment scores. However, most other regions lag behind, indicating spatial disparities in infrastructure quality and visitor satisfaction. These regional differences suggest that the benefits of recent investments have not been distributed evenly across the country.

The distribution of topic prevalence over time reveals additional shifts in tourist focus. While topics such as urban parks, culture and history, and religious sites dominated in earlier years, there has been a clear transition toward tourism services and nature-related experiences since 2018. This shift may reflect evolving visitor preferences as well as the impact of changing government priorities and global tourism trends.

Overall, the analysis of sentiment by topic and region identifies transportation infrastructure as a critical area requiring sustained attention, while also highlighting the continuing success of Kazakhstan's natural attractions in generating positive visitor experiences.

6.4 Kazakhstani Bias in Review Sentiment

The analysis of potential bias among Kazakhstani reviewers, compared to international tourists, offers valuable insights into the reliability and objectivity of user-generated content on platforms like TripAdvisor. The result suggests that, in practical terms, local and international reviewers assess Kazakhstan's attractions with similar levels of sentiment.

The lack of substantial bias from Kazakhstani reviewers indicates that sentiment analyses based on TripAdvisor data can be considered robust and not substantially skewed by national affiliation. This finding supports the broader validity of using online review platforms for comparative and trend analyses in tourism research. It also suggests that the observed patterns of sentiment are shaped by the actual tourism experience rather than by patriotic bias or promotional intent among local reviewers.

6.5 Seasonality Effects in Tourist Sentiment

The results of the time series decomposition and statistical testing demonstrate clear and significant seasonality in tourist sentiment toward Kazakhstan's attractions. The identification of a consistent annual cycle, with sentiment scores peaking in November and reaching their lowest point in March, indicates that visitor experiences and perceptions fluctuate predictably throughout the year. The Kruskal-Wallis test confirmed that these differences in monthly sentiment are statistically significant, supporting the conclusion that recurring seasonal factors play a substantive role in shaping tourist sentiment.

The upward long-term trend in sentiment scores, alongside stable seasonal fluctuations, suggests that improvements in the tourism sector have been sustained over time, even as certain months continue to present challenges for visitor satisfaction. The regularity and strength of the seasonal component imply that environmental, operational, or event-driven variables unique to specific months are consistently impacting the sentiment expressed in reviews.

These findings highlight the need to consider both long-term development and seasonal variability in tourism management and marketing strategies. Addressing the factors contributing to lower sentiment during specific months, such as March, may offer opportunities for targeted improvements in visitor experience and perception during off-peak periods.

6.6 Limitations

Despite the valuable insights provided by this study, several limitations should be acknowledged. First, the exclusive reliance on TripAdvisor reviews may introduce self-selection bias, as those who choose to leave reviews online may not represent the broader population of tourists visiting Kazakhstan. This potentially skews sentiment toward more digitally active, engaged, or opinionated individuals, and may underrepresent visitors who do not use online platforms or who speak less common languages. Additionally, although multilingual sentiment analysis and machine translation tools were employed to process a diverse dataset, some nuances, idiomatic expressions, or cultural meanings may be lost or misinterpreted, especially in languages with fewer resources such as Kazakh language. The dataset's coverage is also limited to the attractions and regions popular among TripAdvisor users, possibly overlooking important but less-visited sites. Additionally, although advanced transformer models significantly enhance analytical capabilities, they may still reflect and propagate underlying biases present in the training data. These biases are not always obvious and can subtly influence sentiment analysis outcomes, highlighting the need for careful scrutiny and responsible use of model predictions. Finally, the analysis is observational and cannot definitively establish causality between observed sentiment trends and policy interventions or external events.

Chapter 7

Conclusion

This research offers a comprehensive and data-driven exploration of tourist perceptions and experiences within Kazakhstan's tourism sector by analyzing over 23,000 multilingual TripAdvisor reviews spanning 2011 to 2024. Leveraging advanced natural language processing and machine learning techniques including BERT-based sentiment analysis, Non-negative Matrix Factorization (NMF) for topic modeling, and zero-shot classification, this research uncovers both broad trends and nuanced insights into how visitors engage with Kazakhstan's attractions, services, and infrastructure.

The primary contribution of this study is its demonstration that user-generated digital reviews can serve as a reliable proxy for real-world tourism activity in Kazakhstan. The strong positive correlation identified between TripAdvisor review counts and official international tourist arrivals validates the utility of digital trace data for supplementing or even partially substituting official tourism statistics, particularly in settings where traditional data may be delayed or incomplete. This finding aligns with broader trends in tourism analytics and offers a foundation for more agile, responsive policy and marketing strategies.

Sentiment analysis reveals an encouraging, upward trajectory in visitor satisfaction over the study period, with a notable acceleration of positive sentiment following 2020. This shift is likely associated with recent improvements in tourism infrastructure and services, supported by targeted government initiative the State Program for the Development of the Tourism Industry (2019–2025). Importantly, this positive trend is not evenly distributed across all regions or experience categories: while natural attractions and hiking consistently elicit the highest sentiment, transportation-related experiences emerge as a persistent pain point, with reviews citing logistical challenges and infrastructure gaps, especially outside Kazakhstan's more developed southern region.

The research also addresses the potential for local bias in digital review data, finding no meaningful difference in sentiment between reviews written by Kazakhstanis and those authored by international tourists. This result bolsters the credibility of TripAdvisor data for objective tourism analytics and suggests that local patriotism or promotional intent does not significantly skew online perceptions of Kazakhstan's attractions.

Finally, this thesis identifies a clear seasonal pattern in tourist sentiment, with satisfaction peaking in November and reaching a low in March. These fluctuations highlight the importance of seasonally targeted marketing and service improvements to smooth out visitor experiences throughout the year.

By systematically combining sentiment analysis, topic modeling, and robust statistical testing, this research fills a significant gap in the academic literature on Kazakhstan's tourism sector. The findings provide actionable insights for policymakers, destination marketers, and industry stakeholders seeking to enhance service quality, address infrastructural weaknesses, and leverage Kazakhstan's unique natural and cultural assets. More broadly, this thesis demonstrates the power of digital review analytics in generating timely, nuanced, and scalable insights into evolving tourism trends, an approach that can be replicated across other emerging destinations.

Future research might extend this work by incorporating additional data sources, such as social media posts or booking platform reviews, and by applying more granular geographic or demographic segmentation to better tailor policy interventions. Continued improvements in machine translation and multilingual sentiment analysis will further enhance the richness and reliability of digital tourism analytics, ensuring that the voices of diverse visitors inform the sustainable growth of Kazakhstan's tourism industry.

Bibliography

[1] K. Charfaoui and S. Mussard, “Sentiment analysis for tourism insights: A machine learning approach,” *Stats*, 2024.

[2] I. Nawawi, K. F. Ilmawan, M. R. Maarif, and M. Syafrudin, “Exploring tourist experience through online reviews using aspect-based sentiment analysis with zero-shot learning for hospitality service enhancement,” *Information*, 2024.

[3] A. Valdivia, M. V. Luzón, and F. Herrera, “Sentiment analysis in tripadvisor,” *IEEE Intelligent Systems*, 2017.

[4] M. C. Enache, “Sentiment analysis in tourism,” *Annals of Dunarea de Jos University of Galati, Fascicle I: Economics and Applied Informatics*, 2020.

[5] O. A. George and C. M. Q. Ramos, “Sentiment analysis applied to tourism: Exploring tourist-generated content in the case of a wellness tourism destination,” *International Journal of Spa and Wellness*, 2024.

[6] J. Pineda-Jaramillo, M. Fazio, M. Le Pira, N. Giuffrida, G. Inturri, F. Viti, and M. Ignaccolo, “A sentiment analysis approach to investigate tourist satisfaction towards transport systems: The case of mount etna,” *Transportation Research Procedia*, 2023.

[7] Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan, “Tourism satellite account of the republic of kazakhstan (2023),” 2023.

[8] K. Pearson, “Note on regression and inheritance in the case of two parents,” *Proceedings of the Royal Society of London*, 1895.

[9] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados, “XLM-T: Multilingual language models in twitter for sentiment analysis and beyond,” *arXiv preprint arXiv:2104.12250*, 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[11] A. R. Pathak, M. Pandey, and S. Rautaray, “Topic-level sentiment analysis of social media data using deep learning,” *Applied Soft Computing*, 2021.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big? ,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[13] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of “bias” in nlp,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[14] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, “The woman worked as a babysitter: On biases in language generation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[15] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, “No language left behind: Scaling human-centered machine translation,” *arXiv preprint arXiv:2207.04672*, 2022.

[16] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*, 2015.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2020. arXiv:1911.02116.

[18] A. A. Lew, “A framework of tourist attraction research,” *Annals of Tourism Research*, 1987.

[19] N. Leiper, “Tourist attraction systems,” *Annals of Tourism Research*, 1990.

[20] Academia.edu, “Tourism development in kazakhstan: Issues and ways forward,” 2023. https://www.academia.edu/106320403/Tourism_Development_in_Kazakhstan_Issues_and_Ways_Forward.

[21] W. Bank, “How new roads in kazakhstan create jobs, save lives, and boost trade,” 2024. https://projects.worldbank.org/en/results/2024/07/11/how-new-roads-in-kazakhstan-create-jobs-save-lives-and-boost-trade.

[22] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, 2015.

[23] U. Gunter and I. Önder, "Forecasting city arrivals with google analytics," *Annals of Tourism Research*, 2016.