Sentiment Analysis of Disneyland Park Reviews

Sophia Wang

Rady School of Management, University of California, San Diego
La Jolla, CA 92093, USA
Email: sophia880221@gmail.com

Xiaoran Wan

Rady School of Management, University of California, San Diego
La Jolla, CA 92093, USA
Email: raewan1254@gmail.com

Abstract

User reviews provide valuable insights into visitor experiences, highlighting factors like attractions, service quality, and wait times. This study analyzes reviews from Disneyland parks in Paris, California, and Hong Kong to examine overall ratings, key influences, and recurring themes in negative feedback.

We apply statistical analysis, machine learning, and deep learning methods, including Logistic Regression, Naïve Bayes, SVM, CNN, and TextCNN, for sentiment classification. Additionally, TF-IDF and Word Vectorization (CountVectorizer) are used to explore relationships between ratings, visitor nationality, and temporal trends.

Findings reveal significant rating differences among the three parks, with key factors including attraction experience, queue times, and service quality. Paris Disneyland shows notably lower ratings, often linked to long wait times and service concerns. This study provides data-driven insights for Disneyland management to improve visitor satisfaction and enhance park operations.

1 Introduction

User reviews are crucial for understanding customer satisfaction in the service industry, particularly in high-profile destinations like Disneyland. In this study, we leverage a dataset from Kaggle that aggregates user reviews from various online platforms, including TripAdvisor and Yelp (Chillar, 2019). This comprehensive dataset covers reviews from Disneyland parks in California, Hong Kong, and Paris. We examine whether deep learning methods, specifically CNN and TextCNN, can provide enhanced sentiment analysis over traditional machine learning approaches in classifying the sentiment of these reviews. Furthermore, we explore the hypothesis that differences in review content across these parks reflect distinct visitor experiences.

2 Dataset

Before diving into the data sources, it is important to note that our dataset offers a comprehensive view of visitor experiences at Disneyland parks (Chillar, 2019). It contains both quantitative ratings and qualitative review texts, allowing us to conduct an in-depth sentiment analysis. The dataset was meticulously cleaned and preprocessed to ensure high-quality, making it a robust foundation for understanding customer satisfaction and identifying key areas for improvement.

2.1 Data Sources

Our study utilizes the *Disneyland TripAdvisor Review Dataset*, which contains visitor reviews and ratings for Disneyland parks in California, Hong Kong, and Paris. The dataset comprises 42,656 records with the following key attributes (Chillar, 2019):

- Review_ID: Unique identifier for each review
- Rating: User rating (1-5 scale)
- Year_Month: Date of the review (Year-Month format)
- Reviewer_Location: Visitor's country of origin
- Review_Text: The textual content of the review
- Branch: Disneyland Park location

The dataset comprises 17,745 reviews for Disneyland California, 11,547 for Disneyland Paris, and 8,255 for Disneyland Hong Kong. Most reviews are positive (79.6% are 4 or 5 stars) compared to 9.7% that are negative (ratings 1 or 2) (Chillar, 2019). To prepare the data for sentiment analysis, duplicates, and missing values were removed, text was converted to lowercase and stripped of punctuation, and temporal features were extracted from the Year_Month field. Reviews with a neutral rating (3 stars) were excluded, and sentiment labels were assigned (positive for ratings 4-5 and negative for ratings 1-2), resulting in a final dataset of 37,547 records.

2.2 Exploratory Data Analysis (EDA)

We perform Exploratory Data Analysis (EDA) to gain an initial understanding of patterns and insights within the Disneyland review data. Through various visualization techniques and statistical analyses, we aim to identify significant differences across parks, highlight visitor sentiment variations by country, and extract common themes in negative feedback to inform subsequent modeling and recommendations.

A. Disneyland Park Rating Distribution

To understand how ratings vary across Disneyland parks, we visualized the average rating per park.

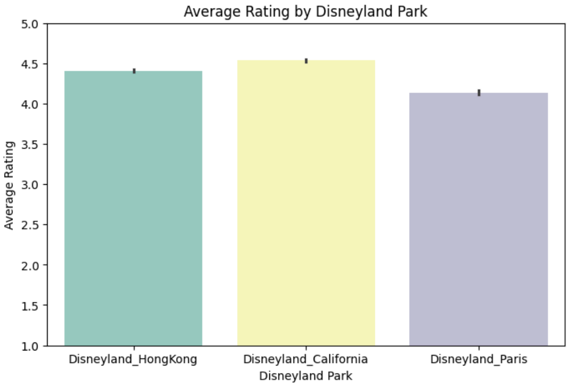

Figure 1: Average Rating by Disneyland Park

California Disneyland has the highest average rating (4.54), followed by Hong Kong Disneyland (4.40). Disneyland Paris has the lowest average rating (4.13), indicating a relatively lower visitor satisfaction compared to the other two parks. A one-way ANOVA test was conducted to examine whether the differences in ratings across the three parks were statistically significant. The results show a highly significant difference ($p < 0.001$), confirming that visitor experiences vary significantly among the parks.

B. Low Rating Analysis by Visitor Country

To examine whether visitor nationality affects ratings, we analyzed the proportion of low ratings (1 & 2 stars) per country.

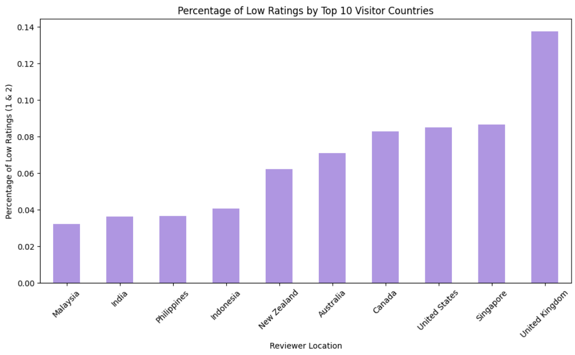

Figure 2: Low Ratings by Top 10 Visitor Countries

Visitors from the United Kingdom (14%) gave the highest proportion of low ratings, followed by Singapore and the United States (~9%). Conversely, visitors from Malaysia, India, and the Philippines had the lowest proportion of low ratings (~3-4%), suggesting a generally more favorable perception of the Disneyland experience. This finding suggests that cultural expectations and service standards in different regions may influence visitor satisfaction.

C. Negative Review Analysis

To identify frequent concerns in negative reviews, we generated a word cloud from reviews labeled as negative.

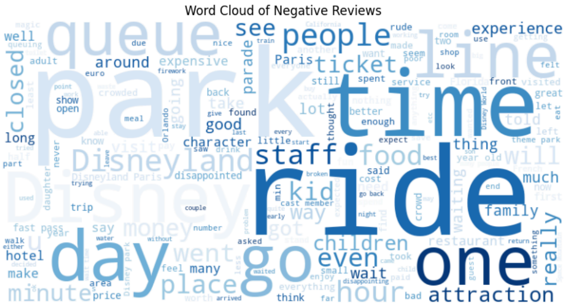

Figure 3: Word Cloud of Negative Reviews

Negative reviews consistently highlight issues with long waiting times ("queue" and "wait"), ride availability and maintenance ("ride" and "attraction"), employee behavior ("staff" and "service"), and high prices ("expensive" and "money"). These insights indicate that improvements in queue management, ride maintenance, service quality, and pricing could significantly enhance the visitor experience at Disneyland parks.

3 Approaches

In this section, we introduce the feature engineering methods used to process the Disneyland review texts, which support the subsequent training of sentiment analysis models. To enhance the performance of our text classification models, we combined several text feature extraction techniques, including TF-IDF, CountVectorizer, and Word2Vec.

3.1 Text Preprocessing

Before building the text classification models, we cleaned and standardized the raw data to ensure consistency and usability. The main preprocessing steps include:

A. Text Cleaning (Clean Review)

This process involves removing punctuation, converting text to lowercase, and eliminating extra spaces and numbers to normalize the text. For example, an original review:

"Thanks God it wasn't too hot or too humid when we visited the park. Otherwise, it would have been unbearable!" (Chillar, 2019)

becomes:

"thanks god it wasn t too hot or too humid when we visited the park otherwise it would have been unbearable."

B. Stopword Removal & Stemming (Processed Review)

We removed common stopwords (e.g., "the", "is", "and") and applied to stem to reduce words to their root forms, thereby reducing noise and retaining terms critical for sentiment classification. For instance: "thank god hot humid visit park otherwise would unbear"
These steps help to minimize noise and improve the model's ability to capture the semantic meaning of the text.

3.2 Feature Extraction

For the textual data, we adopted three complementary feature extraction methods to provide suitable inputs for various types of models. First, we employed TF-IDF (Term Frequency - Inverse Document Frequency) to capture the importance of each word within the documents. This method, widely used in traditional machine learning models such as Logistic Regression, Naïve Bayes, and SVM, selects the top 5000 significant unigrams and bigrams to represent the text. By effectively filtering out overly common words, TF-IDF highlights the terms that are most informative for classification. For example, some of the extracted keywords include "abil", "abl", "abl enjoy", "abl get", "abl go", "abl ride", and "abl see".

In addition, we utilized CountVectorizer to generate word frequency statistics from the text. This method is particularly beneficial for probability-based classifiers like Naïve Bayes and SVM, as it captures the frequency of unigrams and bigrams in each document. The Count Vectorizer approach helps the model to account for the variation in document lengths by emphasizing the most frequently occurring terms. Similar to TF-IDF, the frequent keywords obtained include "abil", "abl", "abl enjoy", "abl get", "abl go", "abl ride", and "abl see", which underscore the core topics present in the reviews.

Finally, to support deep learning models such as CNN and TextCNN, we trained a 100-dimensional word embedding using Word2Vec. This method learns semantic relationships between words by mapping them into a continuous vector space. As a result, words with similar meanings are located close together, enabling deep learning models to capture richer contextual information. For instance, the Word2Vec vector for the term "disneyland" is represented by a numerical vector such as [-1.6249814, -0.17883344, 0.4471472, 0.5960463, 0.17766759, -0.5706056, …]. This semantic representation enhances the model's ability to understand and analyze the nuanced sentiments expressed in the reviews.

3.3 Data Preview

After cleaning and feature extraction, the processed dataset includes the following fields :

Field	Description
Rating	User rating (1-5)
Reviewer_Location	Country of the reviewer
Review_Text	Original review text
Branch	Disneyland park location (California, Paris, Hong Kong)
Sentiment	Sentiment label (1 for positive, 0 for negative)
Clean_Review	Cleaned text after preprocessing
Processed_Review	Text after stopword removal & stemming

Table 1: Overview of Dataset Attributes

Field	Description
Rating	4
Reviewer_Location	Australia
Review_Text	If you've ever been to Disneyland...
Branch	Disneyland_HongKong
Sentiment	1
Clean_Review	if youve ever been to disneyland anywhere...
Processed_Review	youv ever disneyland anywher youll find...

Table 2: Data Sample

4 Experiments

In this section, we present the sentiment classification experiments conducted using both traditional machine learning models and deep learning models. The goal is to evaluate and compare different models in identifying sentiment polarity from Disneyland visitor reviews.

4.1 Model Selection

For sentiment classification, we implemented two types of models: traditional machine learning models and deep learning models. In the traditional category, we used Logistic Regression as a baseline model due to its efficiency in handling high-dimensional text data. Additionally, we employed Naïve Bayes, a probabilistic classifier that operates under the assumption of word independence, and Support Vector Machine (SVM), which is well-regarded for its robust performance in text classification tasks.

For the deep learning approach, we developed two architectures. The first is a Convolutional Neural Network (CNN), specifically designed for short-text classification by leveraging convolutional layers to extract essential text features. The second is TextCNN, a variant of CNN that employs multiple convolutional filters to capture n-gram level features, thereby providing a deeper understanding of sentiment-related phrases.

To evaluate the performance of these models, we utilized several key metrics. Accuracy measures the proportion of correctly classified instances, while the F1-score, which is the harmonic mean of precision and recall, is particularly useful for handling imbalanced datasets. Additionally, we used the AUROC (Area Under the ROC Curve) to gauge how effectively each model distinguishes between positive and negative reviews, and we analyzed confusion matrices to visually assess the distribution of correct and incorrect predictions for each class.

4.2 Experimental Setup

The dataset was split into 80% training and 20% testing to ensure robust evaluation. Feature extraction was performed using TF-IDF for traditional models and Word2Vec embeddings for deep-learning models. The models were then trained and optimized using hyperparameter tuning to achieve the best classification performance.

For machine learning models, TF-IDF feature vectors were used, and models were trained with optimized parameters.
For deep learning models, processed text was tokenized and converted into padded sequences before being fed into CNN and TextCNN architectures.

4.3 Model Evaluation and Results

In this section, we present the detailed evaluation results of our sentiment classification experiments. Using metrics such as Accuracy, F1-score, AUROC, and confusion matrices, we assess the performance of both traditional machine learning models and deep learning models. The analysis that follows is divided into two parts: first, we examine the results obtained from traditional models, and then we present the outcomes from our deep learning approaches. This structured evaluation enables us to comprehensively compare their strengths and limitations in classifying visitor sentiments.

A.Traditional Machine Learning Models

In our experiments with traditional machine learning models, we evaluated the performance of Logistic Regression, Naïve Bayes, and SVM. Logistic Regression achieved an accuracy of 95.43% with an F1-score of 97.51% and an AUROC of 0.80. Naïve Bayes, while slightly lower, attained an accuracy of 93.50% and an F1-score of 96.52%, with an AUROC of 0.68, which indicates a comparatively weaker ability to distinguish between positive and negative reviews. Among these models, SVM performed the best, with an accuracy of 95.75%, an F1-score of 97.67%, and the highest AUROC of 0.84, demonstrating its strong classification capabilities.

Models	Accuracy	F1-score	AUROC
Logistic Regression	95.43%	97.51%	0.8
Naïve Bayes	93.50%	96.52%	0.68
SVM	95.75%	97.67%	0.84

Table 3: Machine Learning Model Performance

B. Deep Learning Models

The CNN and TextCNN models were trained using Word2Vec embeddings. The CNN model achieved an accuracy of 94.71% and an F1-score of 97.11%, with an AUROC of 0.79, which is slightly lower than the performance of the SVM model. In contrast, TextCNN, which employs multiple convolutional filters to extract more granular text features, obtained an accuracy of 94.23%, an F1-score of 96.83%, and an AUROC of 0.80. This indicates that TextCNN slightly outperforms CNN in distinguishing between positive and negative sentiment.

Models	Accuracy	F1-score	AUROC
CNN	94.71%	97.11%	0.79
TextCNN	94.23%	96.83%	0.8

Table 4: Deep Learning Model Performance

4.4 Model Comparison and Business Implications

When comparing the results, SVM emerges as the best traditional model, outperforming both Logistic Regression and Naïve Bayes in terms of F1-score and AUROC. However, the deep learning models, CNN and TextCNN, demonstrated strong recall capabilities, indicating their potential for detecting subtle variations in sentiment that might otherwise be missed.

From a business perspective, these findings provide valuable insights for improving the customer experience at Disneyland. For instance, the frequent mention of long wait times in negative reviews suggests that optimizing queue management, such as introducing fast-track ticketing options, could significantly enhance customer satisfaction. Additionally, the recurring complaints about staff attitude highlight the need for enhanced customer service training to maintain and improve the brand reputation. Moreover, the reviews often note high ticket and food prices, indicating that implementing flexible pricing strategies or promotional offers could improve the perceived value for visitors.

5 Results & Discussion

In this section, we summarize the key results obtained from our data analysis and sentiment classification experiments. We highlight notable differences in visitor experiences across the Disneyland parks, identify prevalent concerns from visitor feedback, and provide detailed evaluations of various sentiment analysis models. Additionally, we discuss practical business implications derived from our findings to help Disneyland management improve overall visitor satisfaction.

5.1 Key Findings

Our analysis explored differences in visitor ratings across Disneyland parks in California, Hong Kong, and Paris. The statistical tests conducted confirmed that there are significant variations in park ratings. Based on the ANOVA results, Disneyland California had the highest average rating (4.54), followed by Disneyland Hong Kong (4.40), while Disneyland Paris received the lowest average rating (4.13). This indicates that overall visitor satisfaction differs across locations, possibly due to factors such as park size, crowd management strategies, service quality, and attraction availability.

Further t-tests comparing Disneyland California and Disneyland Paris showed a highly significant difference ($p < 0.001$), reinforcing the observation that visitor experiences are not uniform across parks. The substantial gap in ratings suggests that Disneyland Paris faces unique challenges that negatively impact visitor satisfaction. These may include longer queue times, fewer ride options, inconsistent service quality, or maintenance issues. In contrast, Disneyland California's strong performance suggests effective crowd control measures, better-maintained facilities, and a more satisfying overall experience.

In addition to quantitative analysis, we examined negative reviews to understand common visitor complaints. Sentiment analysis and topic modeling revealed several recurring issues that significantly impacted visitor satisfaction. One of the most frequently mentioned concerns was long wait times, with visitors expressing frustration over excessive queue lengths, particularly for popular attractions. This issue was consistently reported across all parks but was most pronounced in Disneyland Paris, where visitors frequently noted extended waiting periods.

Another common complaint was ride availability, as multiple reviews cited dissatisfaction with ride closures and maintenance schedules. Visitors expressed disappointment when key attractions were unavailable during their visit, which negatively affected their overall experience. Staff behavior also emerged as a notable concern, with complaints about customer service and staff attitudes being more prevalent in Disneyland Paris compared to the other two locations. This suggests potential gaps in employee training or cultural differences in service expectations that may need to be addressed.

Additionally, high ticket prices were a major point of dissatisfaction. Many visitors felt that the overall cost, including admission fees and in-park expenses, did not align with the value they received. This sentiment was particularly strong in Disneyland Paris and Hong Kong, where visitors perceived fewer high-quality attractions compared to Disney- land California.

These findings indicate that individual parks require tailored operational improvements to enhance visitor satisfaction. By addressing key issues such as queue management, ride availability, staff training, and pricing strategies, Disneyland can create a more consistent and enjoyable experience across all locations.

5.2 Sentiment Analysis Models

To classify visitor sentiment, we compared traditional machine learning models (Logistic Regression, Naïve Bayes, SVM) with deep learning models (CNN, TextCNN). The goal was to determine which approach was most effective in identifying positive and negative sentiments from review texts.

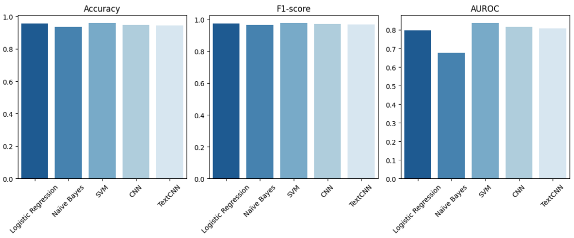

Figure 4: Sentiment Models Performance

SVM achieved the highest accuracy among machine learning models, reaching 95.75%, with Logistic Regression following closely behind at 95.43%. Naïve Bayes, while still performing well, had a slightly lower accuracy of 93.50%. In addition to its high accuracy, SVM also achieved the highest AUROC score of 0.84, demonstrating a strong capability to differentiate between positive and negative reviews with greater precision.

For deep learning models, CNN achieved an accuracy of 94.71% with an F1-score of 97.11%, while TextCNN achieved 94.23% accuracy and an F1-score of 96.83%. TextCNN had a slightly higher AUROC (0.80) compared to CNN (0.79), indicating that its ability to differentiate between sentiments was slightly better.

The results suggest that while SVM remains the most effective traditional machine learning model, deep learning models such as CNN and TextCNN demonstrate strong recall and are particularly useful for capturing nuanced sentiment expressions.

5.3 Business Implications

The insights from sentiment analysis provide actionable recommendations to improve visitor experiences at Disneyland parks. Based on the analysis of ratings and negative reviews, we propose the following three key areas for improvement.

A. Optimizing Queue Management

One of the most frequently mentioned complaints in negative reviews is long wait times. The word cloud analysis highlights words such as "queue," "wait," and "time" as dominant concerns. To mitigate this issue, Disneyland could introduce enhanced queue management systems such as digital wait-time estimations, expanded FastPass access, and improved ride scheduling to reduce visitor frustration and improve the overall experience.

B. Improving Staff Training and Service Quality

Negative reviews frequently mention issues related to staff behavior, with words like "rude" and "staff" appearing prominently. To address this, Disneyland should focus on enhancing customer service training programs to ensure employees are well-equipped to handle guest interactions professionally and courteously. Improved training could lead to better visitor-staff interactions, reducing negative feedback related to service quality.

C. Adjusting Pricing and Value Perception

Many visitors criticized Disneyland's high ticket prices and expensive food, with "expensive" and "food overpriced" frequently appearing in negative reviews. To enhance satisfaction, Disneyland could offer seasonal discounts and family packages while introducing more affordable dining options. These adjustments would improve perceived value and encourage longer stays.

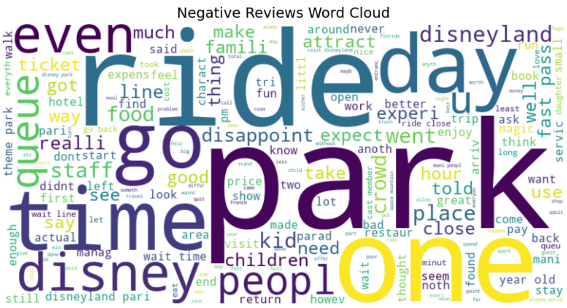

Figure 5: Negative Reviews Word Cloud

6 Conclusion & Future Work

In this final section, we summarize the major insights derived from our sentiment analysis and discuss their implications for enhancing visitor experiences at Disneyland. We also reflect on the effectiveness of our modeling approaches and outline promising directions for future research.

6.1 Conclusion

This study analyzed visitor reviews across different Disneyland parks and applied both machine learning and deep learning models for sentiment classification. Based on the findings, we identified significant differences in ratings among Disneyland parks, with Disneyland California receiving the highest average rating (which is 4.54) and Disneyland Paris the lowest (which is 4.13). The primary factors contributing to lower ratings included long wait times, high ticket prices, staff service quality, and ride maintenance issues.

The sentiment classification models performed reliably, with SVM and Logistic Regression demonstrating superior accuracy and F1-score among traditional machine learning models. Meanwhile, CNN and Text-CNN also achieved competitive classification performance while being more efficient than LSTM in training speed. These results suggest that sentiment analysis can effectively uncover key visitor concerns and provide actionable insights for Disneyland’s management.

Key factors affecting visitor satisfaction include ride maintenance, where temporary ride closures and maintenance issues negatively impact visitor experience, long queue times that lower customer satisfaction, negative feedback on staff service indicating the need for enhanced training, and high ticket and food prices that affect the perceived value and overall sentiment. Further analysis of negative reviews validated these findings, providing Disneyland's management with clear directions for improvement.

6.2 Future Work

While this study provides a comprehensive analysis of visitor sentiment, there are several areas for further improvement and expansion. Currently, our analysis covers Disneyland parks in California, Paris, and Hong Kong; however, future research could incorporate reviews from Shanghai Disneyland to create a more global comparison of visitor experiences.

In addition, although CNN and TextCNN outperformed LSTM and BERT in training efficiency, future research could explore Transformer-based models such as Distil-BERT. These models may further improve classification accuracy and better capture the contextual nuances in reviews, providing deeper semantic understanding and improved generalization.

Moreover, the current study classifies sentiment as simply positive or negative. Future work could incorporate sentiment intensity analysis to distinguish between varying levels of dissatisfaction, such as mild complaints versus severe dissatisfaction. This finer-grained analysis could offer even more detailed insights into visitor concerns.

Overall, our study provides several data-driven recommendations for improving visitor satisfaction at Disneyland parks, which demonstrates the effectiveness of deep learning in sentiment analysis. By integrating advanced NLP techniques and expanding data sources, future research can further enhance sentiment classification accuracy and provide even more precise insights for Disneyland's management.

References

Chillar, Arush. 2019. *Disneyland Reviews: Reviews and Ratings of 3 Disneyland branches - California, Hong Kong, and Paris*. Kaggle. Retrieved March 13th, 2025, from https://www.kaggle.com/datasets/arushchillar/disneyland-reviews/data.

Work Link: https://github.com/sophia0507/Sentiment-Analysis-of-Disneyland-Park-Reviews.git