Disneyland Reviews & NLP

(BERT MODEL & TOPIC MODELING)

Gerry Cruz | STAT 574 | May 9th 2024

Submitted to

Prof. Olga Korosteleva

Report prepared by

Gerry Cruz

Contents

I. Introduction **2**

II. Background **2-3**

III. Data Description **3-4**

IV. Results **4-13**

V. Conclusion **13**

VI. Appendix **14-21**

a) Figure 3(a) 14

b) Figure 4(a) 15

c) Figure 3(b) 16

d) Figure 4(b) 17

e) Figure 3(c) 18

f) Figure 3(d) 19

g) Figure 3(e) 20

h) Figure 3(f) 20

i) Figure 3(g) 21

VII. References **22**

I. Introduction

After being introduced to Statistical models using textual data in Data Mining class, I realized my experience lacked projects involving text data. This prompted my interest in Natural Language Processing (NLP). The allure of deciphering the intricacies of human language and extracting insights from textual information fascinated me.

Recently, during a job fair visit by Disneyland Parks, a potential employer, I envisioned an opportunity to showcase my Data science skills. I aspired to create a project that would not only demonstrate my proficiency but also provide valuable insights to Disneyland Parks, particularly regarding how visitors perceive their park experience.

While looking for datasets on Kaggle.com, I came across a dataset containing text data related to Disneyland Park in California, presenting the perfect opportunity to apply my burgeoning NLP skills. This dataset piqued my curiosity, offering a glimpse into the sentiments and experiences of park visitors. I embarked on this project with enthusiasm, eager to unravel the stories hidden within the textual narratives and unveil valuable insights for Disneyland Parks.

II. Background

Big-name brands like The Walt Disney Company, which owns Disneyland Parks, are known for providing entertainment and unforgettable experiences to their customers. They prioritize delivering exceptional experiences that leave a lasting impression on visitors. It's crucial for companies like Disney to care about their customers' experiences with their products, including the theme park experience. Ensuring that visitors have a positive and memorable time at the park not only pleases the customer but also increases the likelihood of them returning in the future.

One effective way to gauge customer feedback and satisfaction is by analyzing their experiences at the park. This involves examining what people say and feel about their interactions and encounters within the park environment. Methods such as word clouds, sentiment analysis, and topic modeling offer valuable insights into visitor sentiments, overall satisfaction levels, and trending topics of discussion related to their park experience.

By delving into customer feedback and sentiment analysis, theme parks like Disneyland can gain a deeper understanding of visitors' perceptions, preferences, and areas for improvement. This proactive approach allows them to tailor their offerings, enhance customer satisfaction, and cultivate a loyal customer base. Ultimately, prioritizing customer experience not only ensures visitors leave the park happy and motivated to return but also contributes to the long-term success and reputation of the brand.

III. Data Description

The dataset, sourced from Kaggle, contains 42,000 reviews across three Disneyland Park branches. The original columns included Review_ID, Rating, Year_Month, Reviewer_Location, Reviewer_Text, and Disneyland_Branch. However, I excluded the Review_ID column as it wasn't essential for the analysis. The focus was on observing Disney California Park, so I did have to limit my dataset to only having reviews from only the California branch. While working with the BERT model, the computation time was extensive and posed complications due to time constraints. Consequently, I reduced the dataset to 150 reviews for BERT model processing. For all other analyses, I utilized the complete set of 42,000 reviews.

IV. Results(1-Code in Python, 2-Code in R)

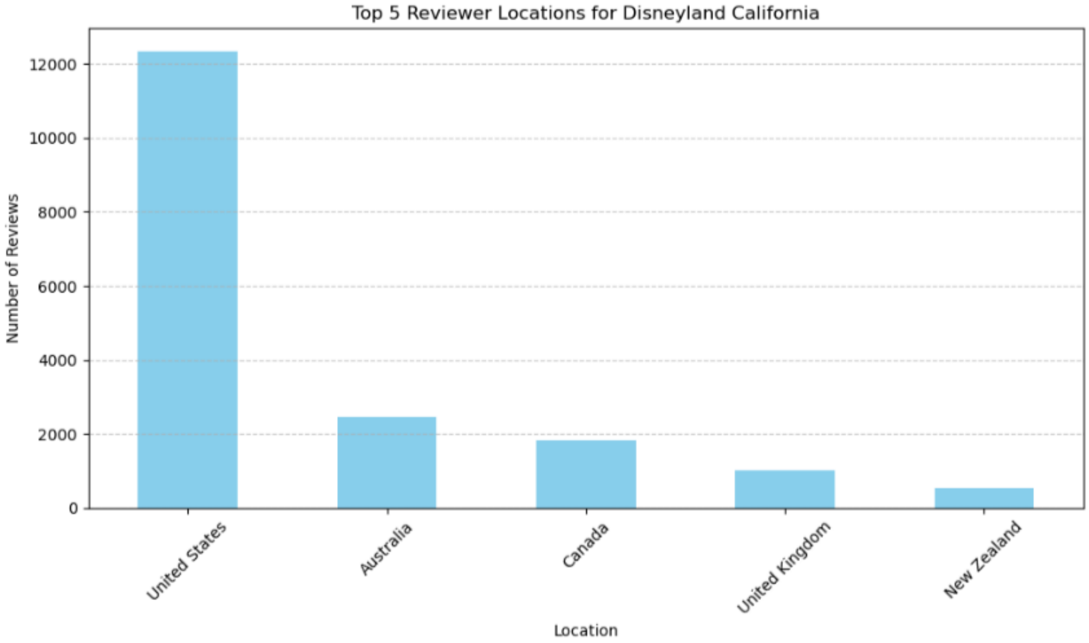


Figure 1(a)

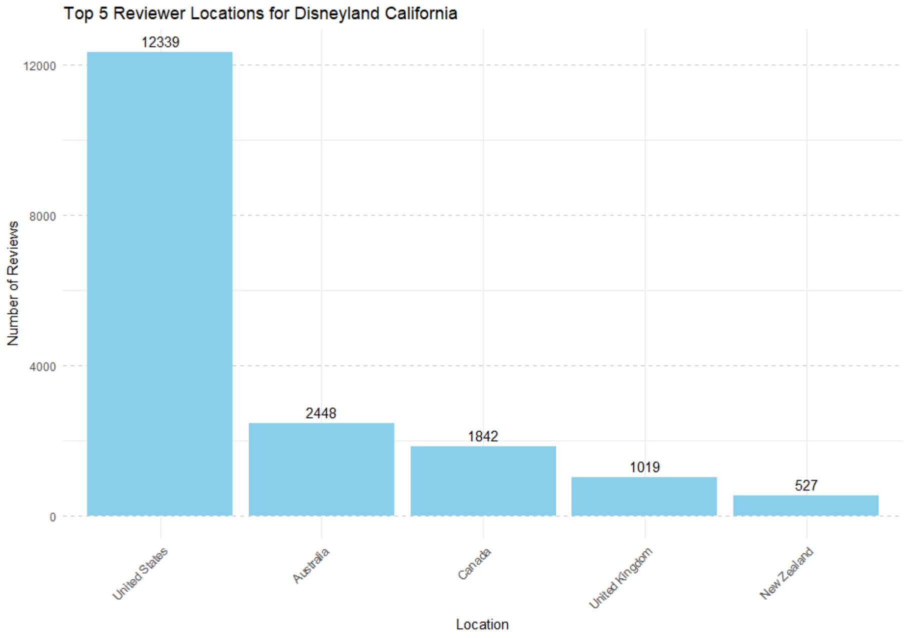


Figure 2(a)

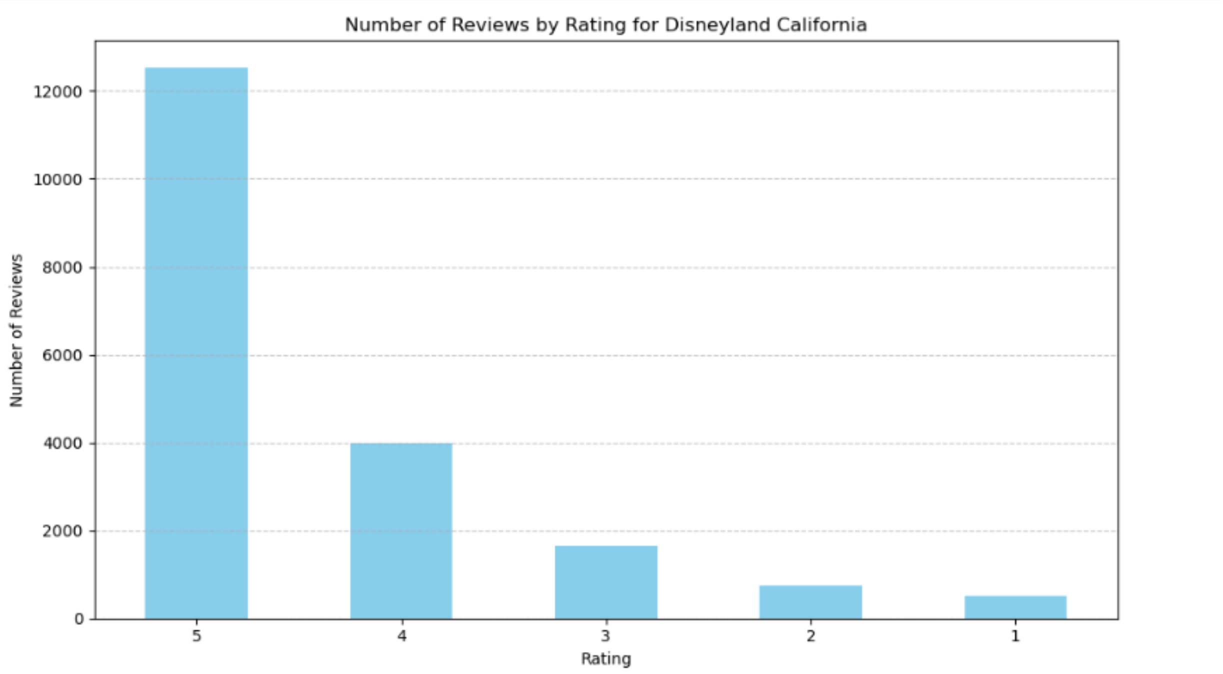


Figure 1(b)

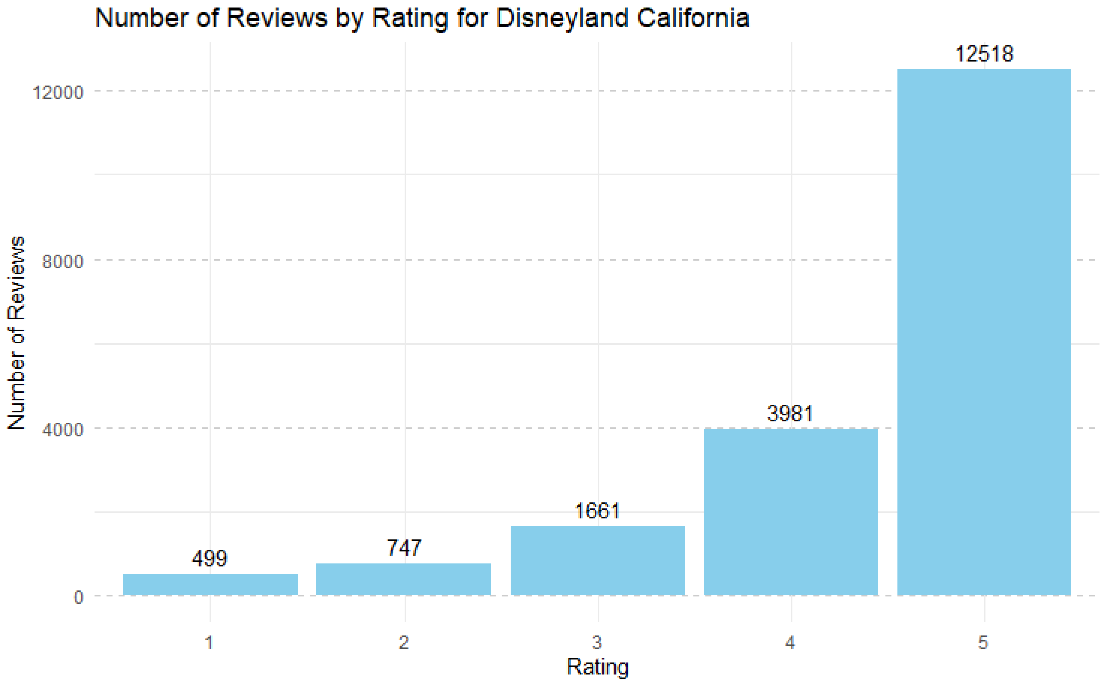
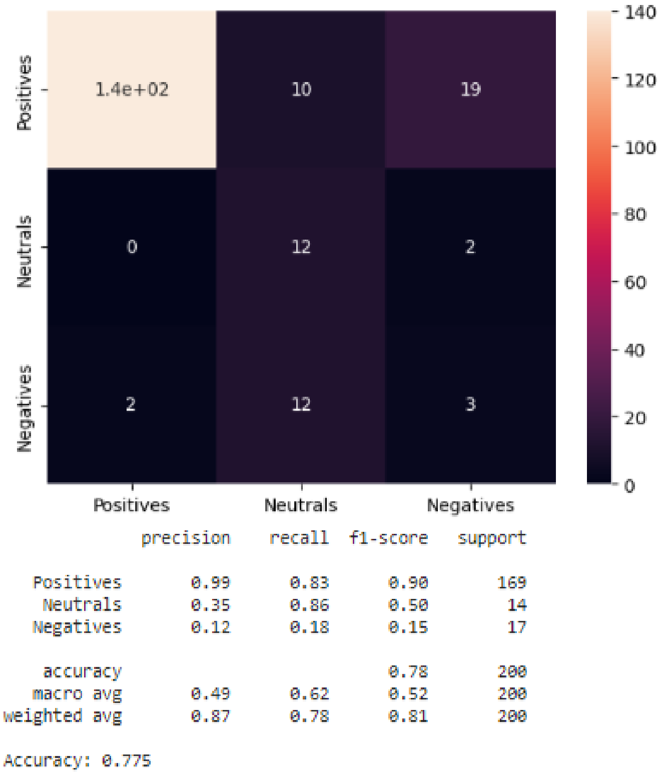

Number of Reviews by Rating for Disneyland California
Number of Reviews
12000
8000
4000
0
499
747
1661
3981
12518
1
2
3
4
5
Rating

Figure 2(b)



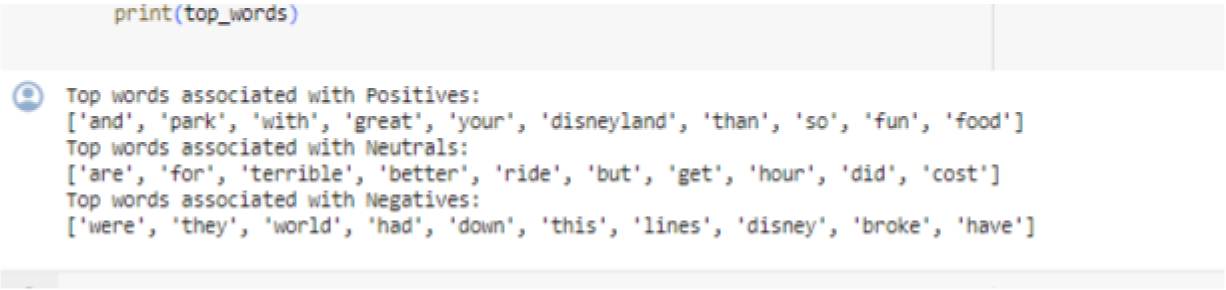
```
              precision    recall  f1-score   support

   Positives       0.99      0.83      0.90       169
    Neutrals       0.35      0.86      0.50        14
   Negatives       0.12      0.18      0.15        17

    accuracy                           0.78       200
   macro avg       0.49      0.62      0.52       200
weighted avg       0.87      0.78      0.81       200

Accuracy: 0.775
```

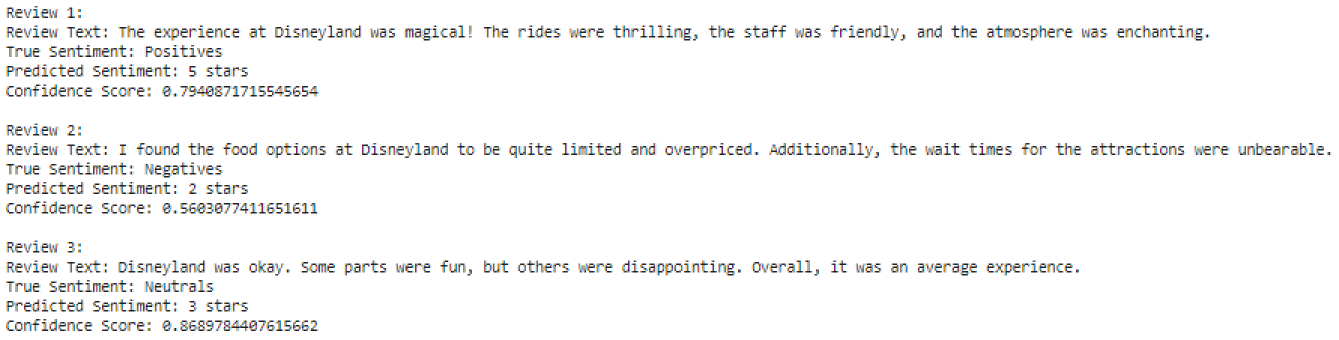
Figure 1(c)



```
print(top_words)
```

```
Top words associated with Positives:
['and', 'park', 'with', 'great', 'your', 'disneyland', 'than', 'so', 'fun', 'food']
Top words associated with Neutrals:
['are', 'for', 'terrible', 'better', 'ride', 'but', 'get', 'hour', 'did', 'cost']
Top words associated with Negatives:
['were', 'they', 'world', 'had', 'down', 'this', 'lines', 'disney', 'broke', 'have']
```

Figure 1(d)



```
Review 1:
Review Text: The experience at Disneyland was magical! The rides were thrilling, the staff was friendly, and the atmosphere was enchanting.
True Sentiment: Positives
Predicted Sentiment: 5 stars
Confidence Score: 0.7940871715545654

Review 2:
Review Text: I found the food options at Disneyland to be quite limited and overpriced. Additionally, the wait times for the attractions were unbearable.
True Sentiment: Negatives
Predicted Sentiment: 2 stars
Confidence Score: 0.5603077411651611

Review 3:
Review Text: Disneyland was okay. Some parts were fun, but others were disappointing. Overall, it was an average experience.
True Sentiment: Neutrals
Predicted Sentiment: 3 stars
Confidence Score: 0.8689784407615662
```

Figure 1(e)

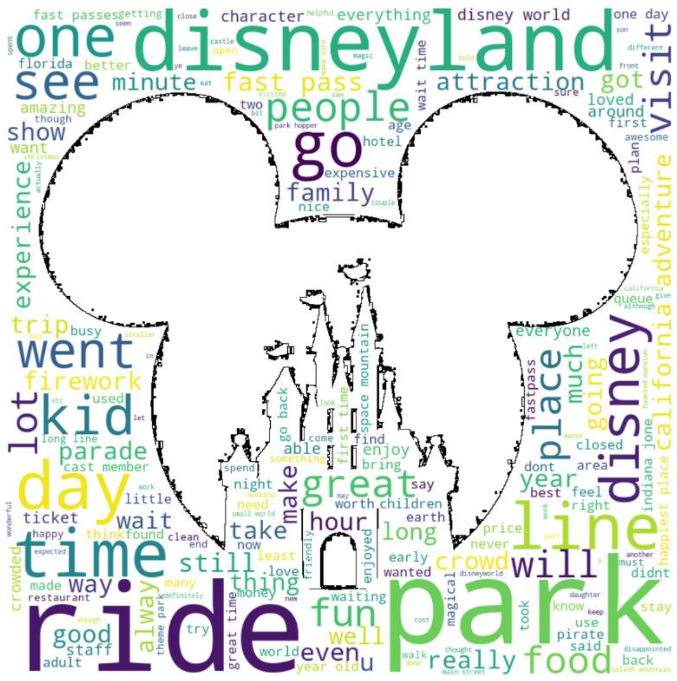


Figure 1(f)

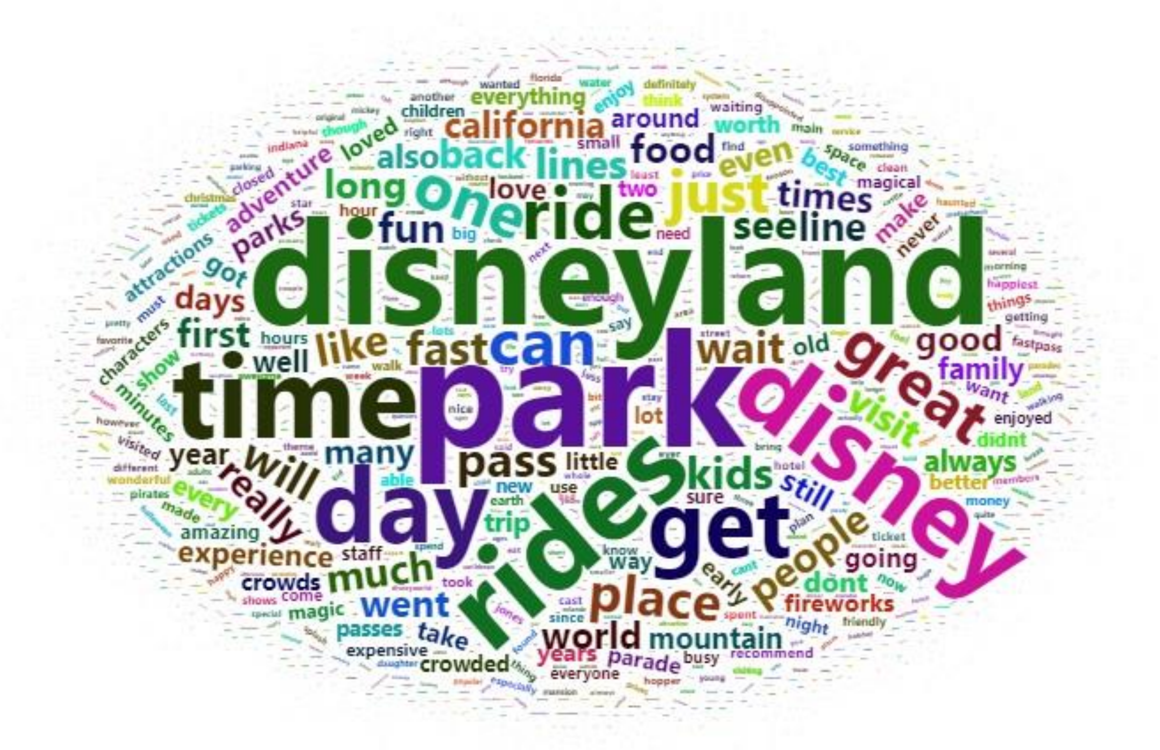


Figure 2(f)

Positive Word Cloud

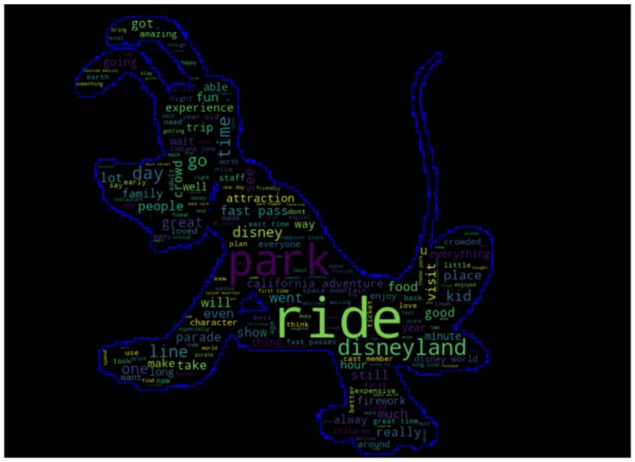


Figure 1(g)

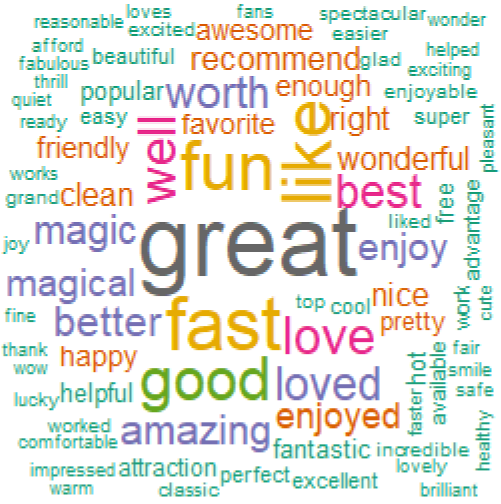


Figure 2(g)

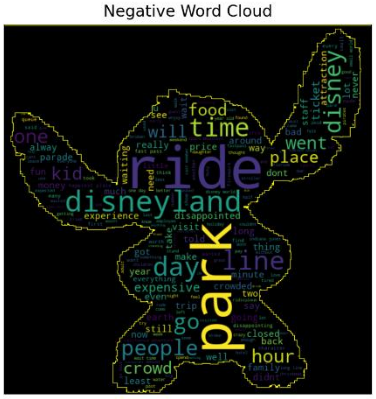

Negative Word Cloud

Figure 1(h)

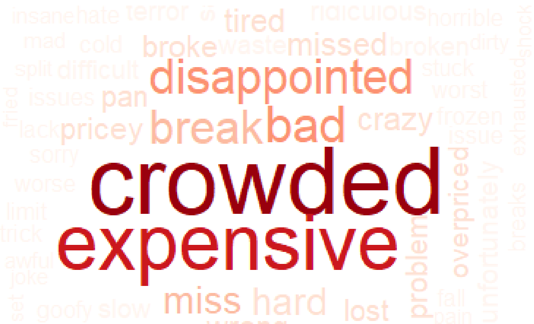
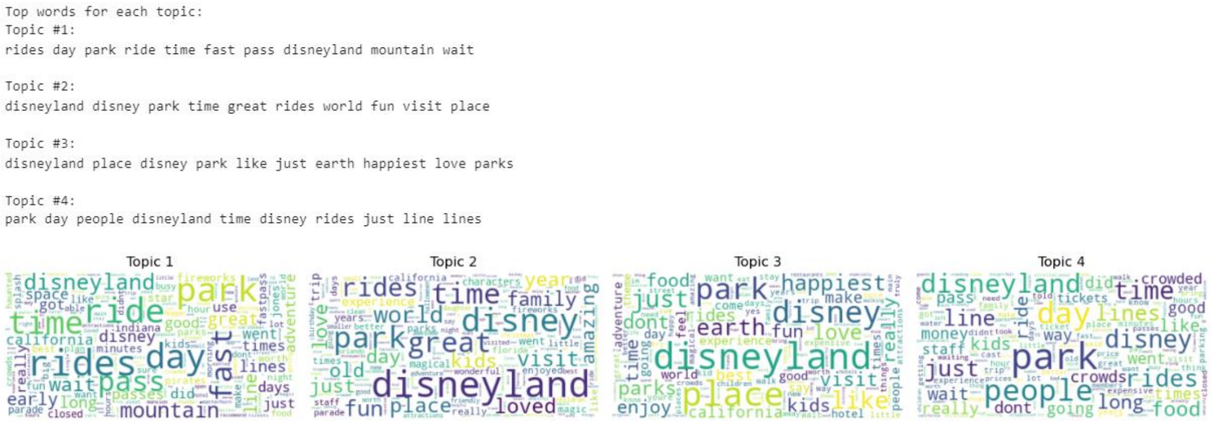


Figure 2(h)



```
Top words for each topic:
Topic #1:
rides day park ride time fast pass disneyland mountain wait

Topic #2:
disneyland disney park time great rides world fun visit place

Topic #3:
disneyland place disney park like just earth happiest love parks

Topic #4:
park day people disneyland time disney rides just line lines
```

Figure 1(i)

After cleaning the data, one of my initial objectives was to identify the top five reviewer locations. Intuitively, I assumed that most reviewers would be from the United States. I created a bar chart (Figure 1/2(a)) to verify this assumption and observed that, as expected, most reviewers were from the United States. Canada was the second largest group, followed by the United Kingdom and New Zealand.

Next, I analyzed the park's ratings. I explored the 'Rating' column and created bar plots (Figure 1/2(b)). The results showed that most reviews rated the park as 'Excellent' (a score of 5), with ratings ranging from 1 to 5.

Following this, I decided to use a BERT model to analyze sentiment in the text reviews and assess its performance on new data. I split the dataset into training and testing sets with an 80/20 split. After training the model, it achieved a prediction score of 0.775 (Figure 1(c)), which was promising. The confusion matrix revealed that the model performed well when predicting positive sentiments.

I then identified the most prominent positive, negative, and neutral words using the BERT model. I used multinomial logistic regression (Figure 1(d)) to find that the top positive words were 'great,' 'Disneyland,' 'fun,' and 'food.' The top negative words were 'world,' 'lines,' and 'broke.' The top neutral words included 'terrible,' 'better,' 'ride,' 'hour,' and 'cost.'

The next step was to see how the model would classify (Figure 1(e)) sentences I provided. I created three sentences, and the model accurately predicted their sentiments as positive, negative, and neutral. I then constructed word clouds to visualize the sentiment expressed toward certain words (Figures 1(f-h)). Despite some overlap in sentiment across different words, I filtered out key terms after analyzing the surrounding words.

To conclude the analysis, I explored the general discussion around visitors' experiences by building a topic modeling algorithm (Figure 1(i)) using LDA (Latent Dirichlet Allocation). I identified five topics, each containing ten words. The overall discussion emphasized the attractions' fun and excitement. Topic number three stood out to me, highlighting the emotional and sentimental connection visitors have with Disneyland, often referred to as 'the happiest place on Earth.' Words like 'love' and 'happiest' reveal the deep affection and joy associated with the park.

V. Conclusion

In conclusion, the analysis of Disneyland Park reviews revealed insightful patterns in visitor feedback. Sentiment analysis using the BERT model and topic modeling with LDA uncovered the positive emotions and enjoyment most visitors experience, as well as key areas for improvement, like long wait times. These findings offer Disneyland Parks valuable insights that can help refine their attractions, enhance visitor satisfaction, and strengthen their reputation as "the happiest place on Earth."