April 22nd 2024

NLP Analysis of Google Reviews for Saudi Arabian Sites

Methodology:

Data transformation: (ratings, tags columns)

- Unpack JSON encoded columns by parsing strings into basic python data structures like lists and dictionaries using literal_eval function from the ast library.
- In case of list of dictionaries, loop through each row to access dictionary keys.
- Pass columns into JSON normalization to normalize semi-structured JSON data into a flat table.

Data transformation: (tags_mapping column)

- Unpack list of place categories and locations into separate columns by accessing list values.

Text cleaning: (content column)

- Use regex to remove Google translation and keep original text only.
- Remove punctuations.
- Normalize special Arabic characters.
- Use lower case (English content).
- Remove longation (Arabic content).
- Remove stopwords.

Sentiment analysis:

There are so many pre-trained models for text classifications. In this case study, I'm implementing transfer learning using BERT model from Hugging face transformers library for multilingual sentiment text classification since we have both Arabic and English reviews.

- Tokenization:
  Converts a string to a sequence of ids (integer), using the tokenizer and vocabulary.
- Use model for BERT text classification on preprocessed content column.

Output will be a score from 1-5.

Where 1 is the most negative sentiment and 5 is the most positive sentiment.

Notes:

1- Multiple BERT model variations were used to test out against the case study text and the model chosen at the end was:

nlptown/bert-base-multilingual-uncased-sentiment.

2- It might be worthwhile to finetune pretrained model to yield better sentiment predictions in the future.

Exploratory Data Analysis:

Hypothesis test:

used to test the existence of correlation between rating and sentiment predictions or lack thereof.

Sentiment and rating categories:

Positive: 5 or 4 score.

Neutral: 3 score.

Negative: 2 or 1 score.

Define null and alternative hypotheses:

H0: rating and sentiment are NOT related.

H1: rating and sentiment ARE related.

P-value: probability of H0 being True.

chi square test was used to test the correlation between two categorical variables (rating_category and sentiment_category)

p-value from chi2 test is less than 0.05, thus we can conclude that the two variables are related.

In other words: p-value which is the probability of H0 being true is very small that it can't be accepted.

April 22nd 2024

Overall distribution of Google review sentiment by source:

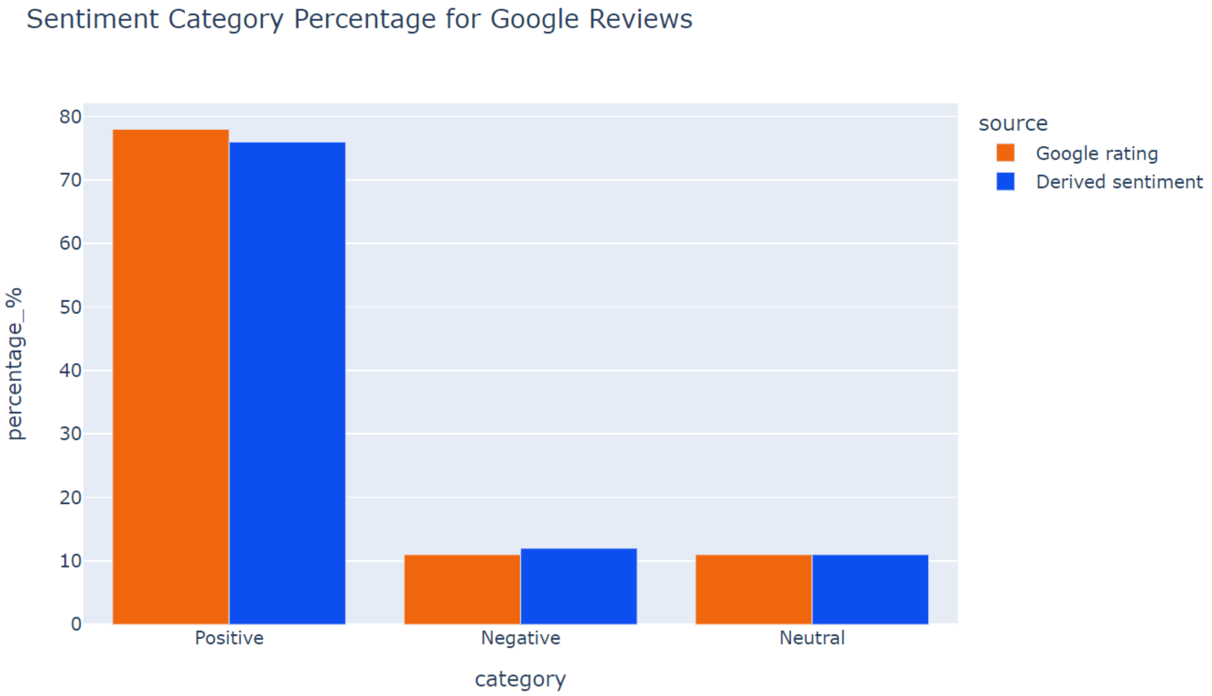

Figure1: Sentiment Category Percentage for Google Reviews by source.

Figure1 above shows that overall, majority of customer feedback is positive with percentages of high 70s for the sample provided. Additionally, Negative feedback percentage is about 12 %.

April 22nd 2024

Overall Google review sentiment frequency per location:

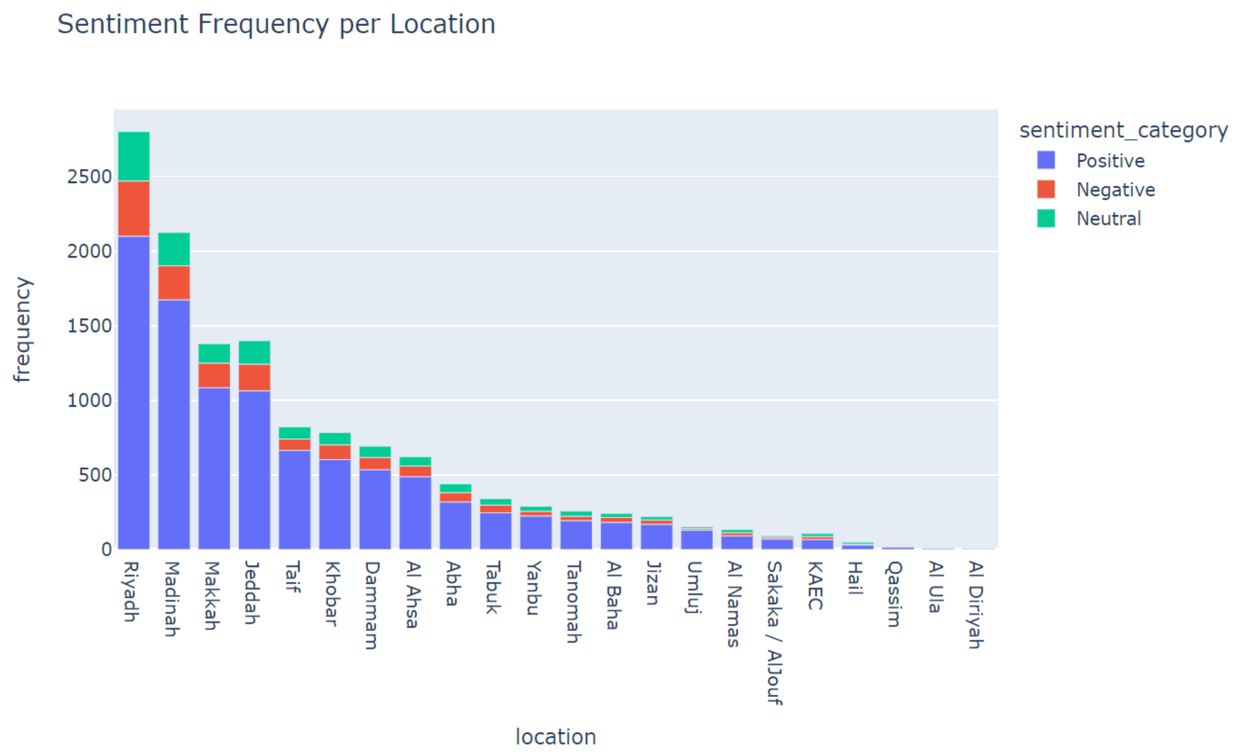

Figure2: Overall Sentiment Frequency per Location.

Note: all bar charts are attached in a separate file of this submission in order to leverage dynamic nature of plotly figures (please feel free to hover over each bar for more information)

Figure2 demonstrates that Riyadh city has the highest count of positive reviews as well as highest count of negative reviews. Followed by Madinah and Makkah cities. On the other hand, Al Diriyah city has the lowest count of positive reviews and no negative reviews in this sample.

April 22nd 2024

Negative sentiment distribution by location and category:

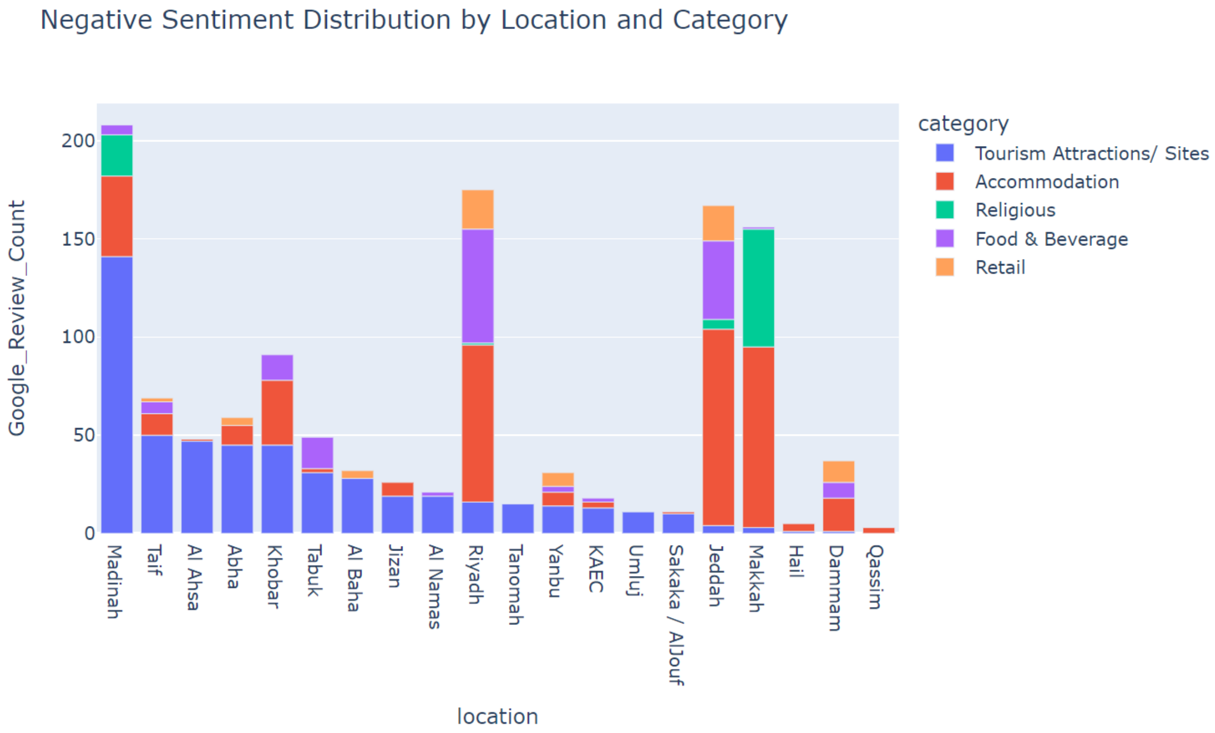

Figure3: Negative Sentiment Distribution by Location and Category.

Figure3 shows that Tourism attractions/sites and accommodations have the most negative review counts across all locations. On the other hand, retail and religious categories reported the lowest review counts.

Most Tourist Attractions with negative reviews:

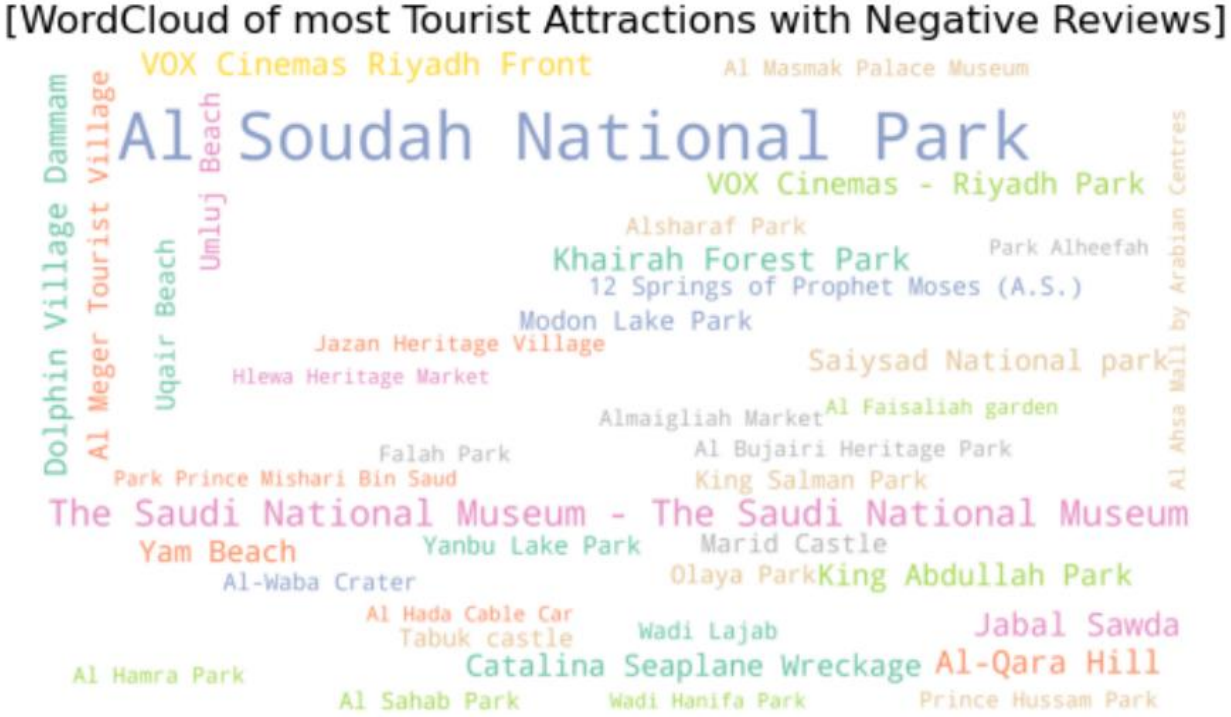

Figure4: Word Cloud of tourist attractions with highest counts of negative reviews.

Figure4 demonstrates places with highest counts of negative reviews in larger font size and places with lowest counts of negative reviews in smaller font size.

Visitor concerns about Al Soudah National Park:

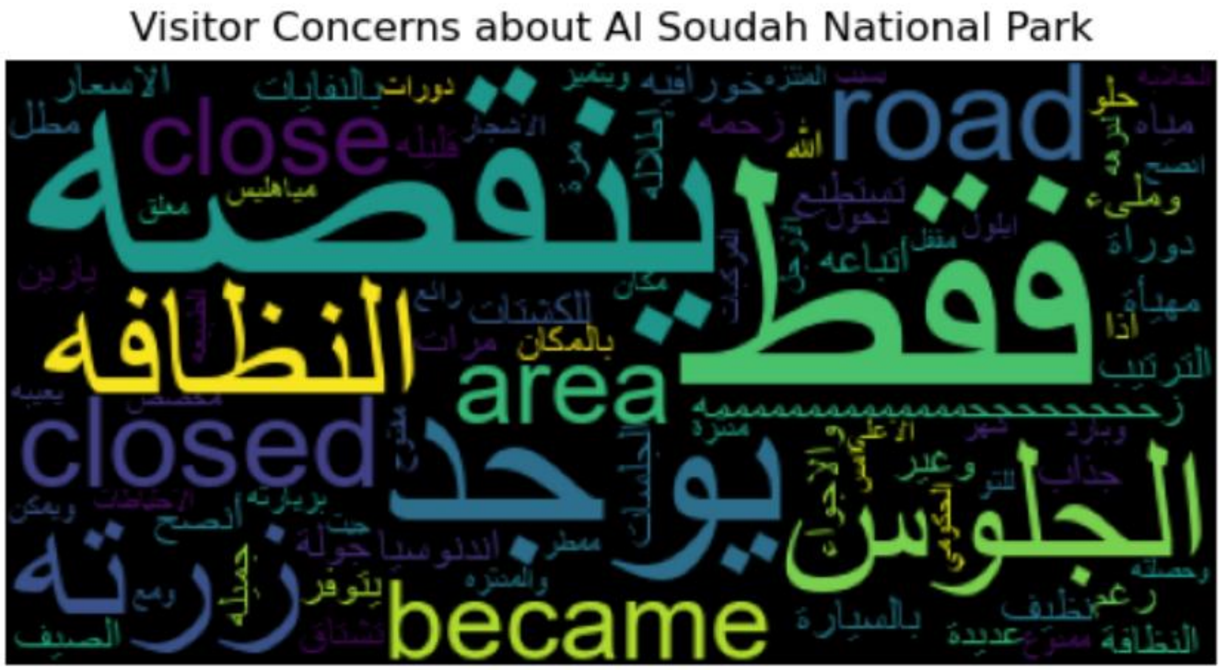

Figure5: Word cloud of Visitor Concerns about Al Soudah National Park.

Figure5 illustrates that most concerns from Al Soudah National park are related to cleanness, trash, restrooms, park closure, road issues.

<u>Recommendations:</u>

Overall, most customer feedback is positive and encouraging which reflects nation wide efforts to deliver a positive experience to customers. In order to reduce negative customer experience, main contributing categories should be focused on: Tourism Attractions and Accommodations.

Further more, each category should be sub-categorized by title column first then study each sub-category concerns to better enhance customer experience.

Recommendation to improve park visitors experience (under tourism category):

1- Improve park cleanness.
2- Improve restroom facilities.
3- Improve trash management.
4- Update park working hours regularly.

April 22nd 2024

Recommendation to improve cinema visitors experience (under accommodation category):

1- Improve service provided.
2- Improve food menu and quality of food.
3- Improve booking experience.