Predicting Hotel Closure in LA

CIS509

Team 480

Table of contents

Introduction

- Background
- Objective
- Data sources

Exploration

- Data cleaning
- Statistics summary
- Data exploration

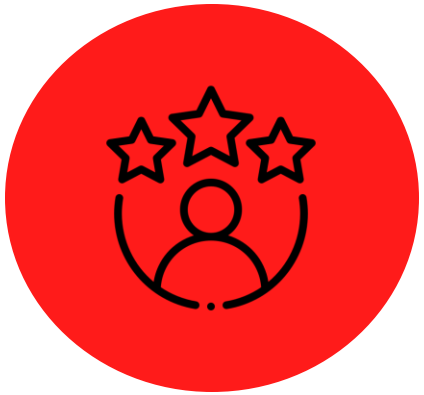
Analysis

- Logistic regression model
- Sentimental analysis
- Topic modelling

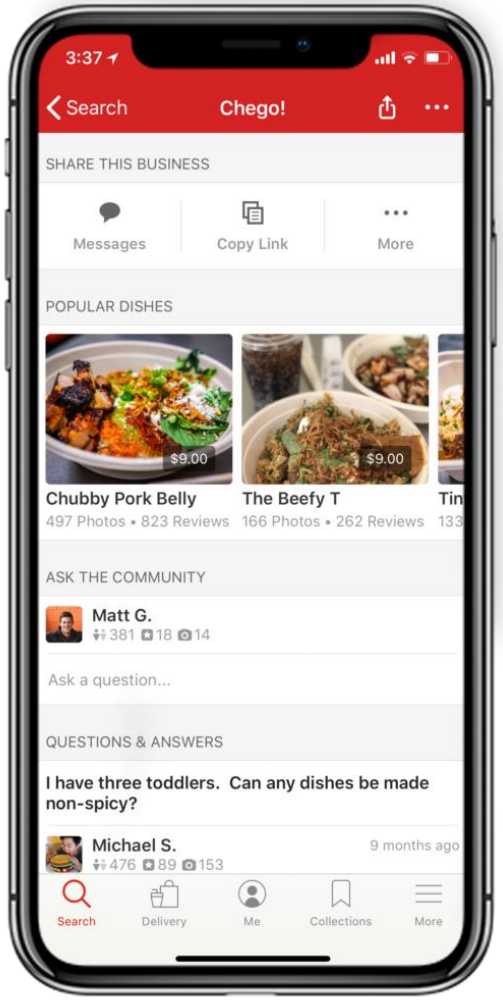
Insights

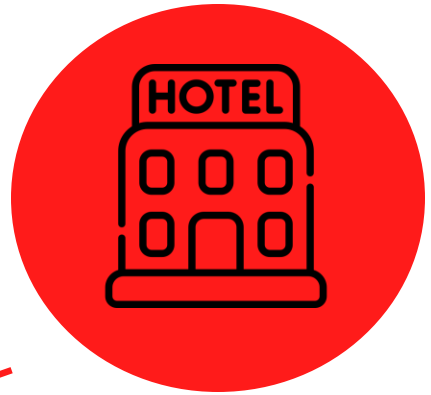
- Conclusion
- Limitations and future considerations

Why are we interested?



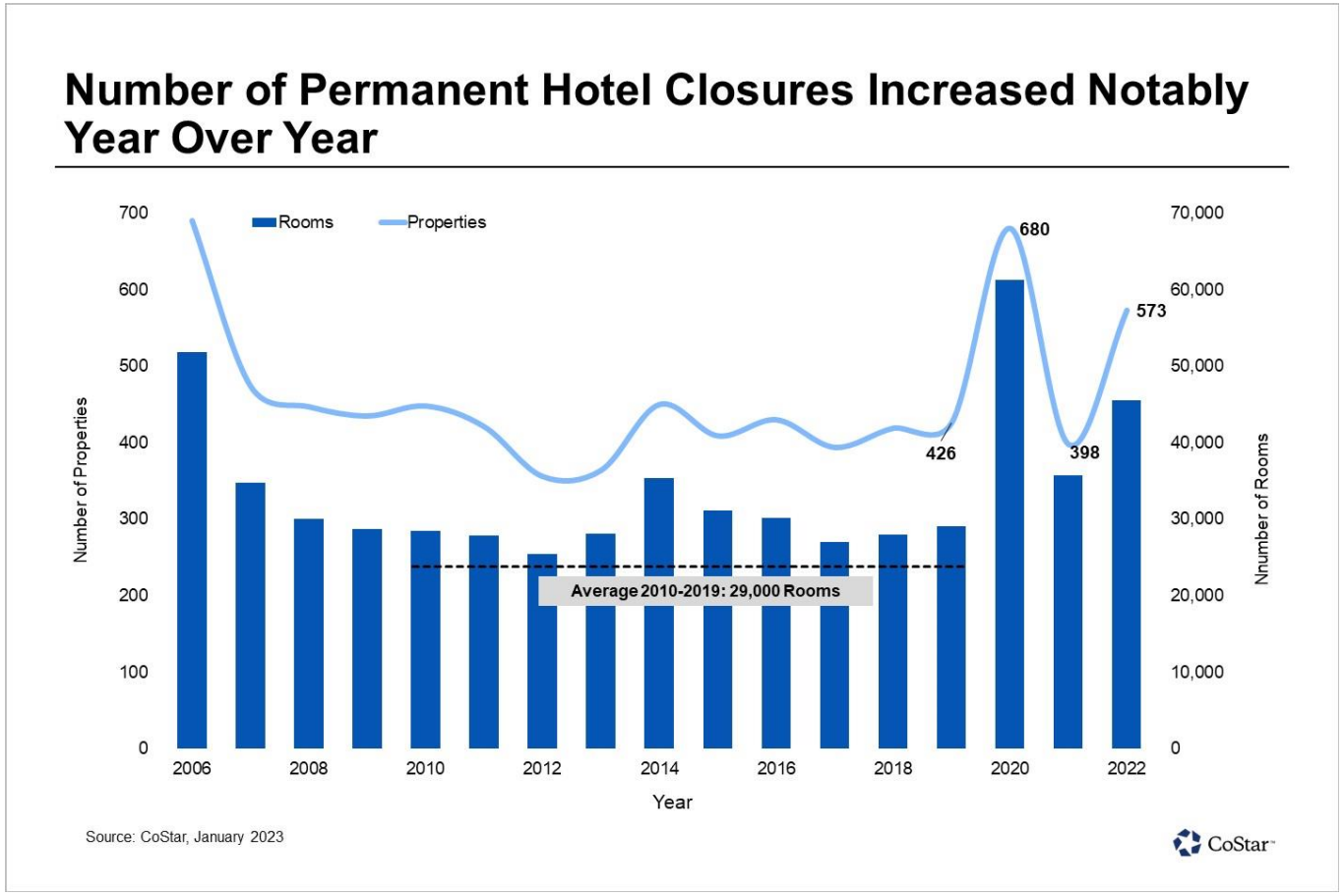
- Customers rely on applications such as Yelp for reviews before trying out new businesses, be it a new hotel, café or a restaurant
- Customers express their level of satisfaction in a business using their words and star ratings





- Higher rating would drive business for the hotels and profit for their investors
- Help them understand their competition and
- Give them an understanding of what customers think about their business

Why is it important?



- Hospitality industry has witnessed a surge in permanent hotel closures.
- Early intervention through predictive analysis allows stakeholders such as investors to intervene with strategic measures
- Analyzing customer experiences and satisfaction may help hotel owners implement targeted improvements
- Empowers stakeholders, including government agencies and tourism boards, to proactively support struggling establishments and implement policies to sustain the hospitality sector's growth

Objectives & data source

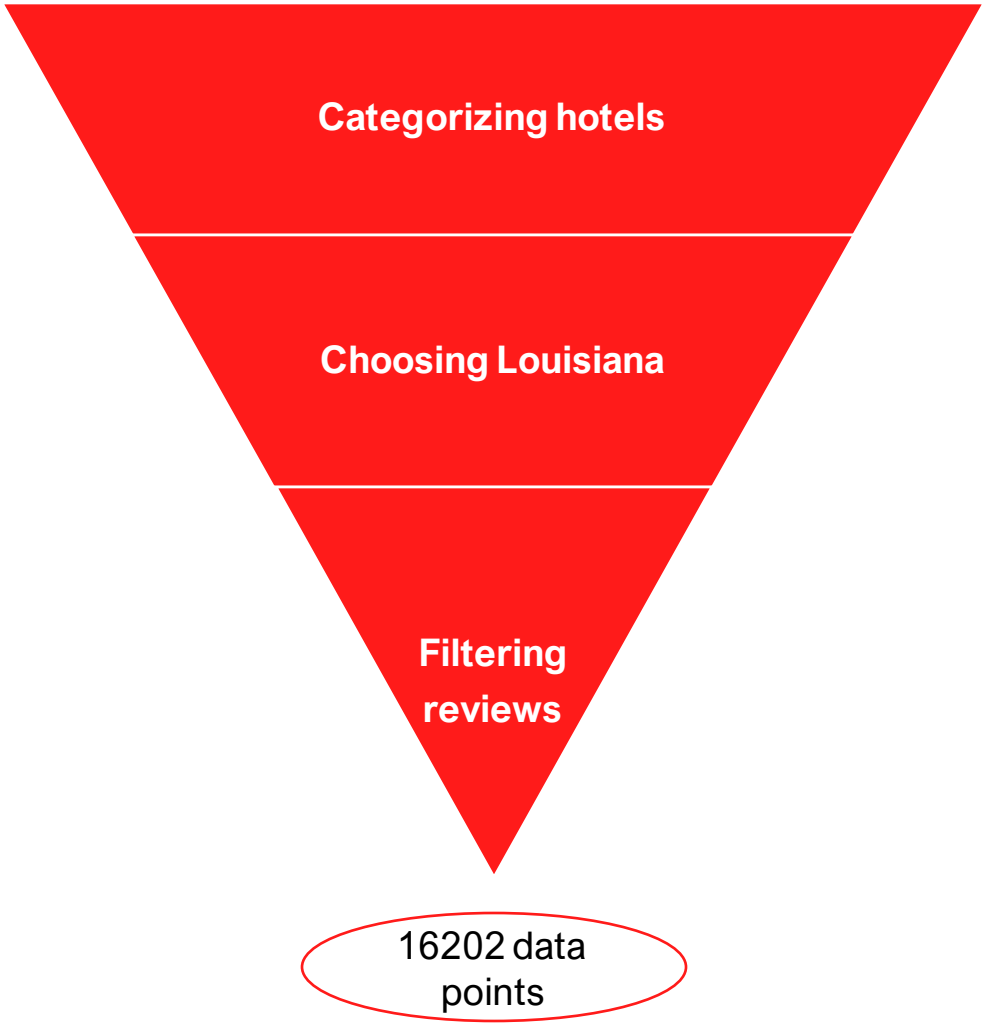
Questions of interest

- Understanding the effect of spatial data (location) and temporal data (dates) on predicting hotel closure
- Exploring the Impact of Customer Sentiments on Hotel Operational Status
- Identifying the key topics and themes present in hotel reviews for both open and closed businesses, and how these topics differ between the two categories

Data Sources

- Yelp Data set on KAGGLE
- **Data Category:** Hotels dataset
- **Scope:** Data spans across businesses categorized as hotels in Louisiana in North America

Data Cleaning



- Extracting business data for business categorized as “**Hotels**”
- Choosing business data for hotels in **“Louisiana”**
- Dropping null values and deleting duplicates
- Filtering short reviews length
- Filtering reviews in other languages ie not in English

Statistical Summary

Total number of hotels

296

Total number of cities

1 (New Orleans, LA)

Average Rating

3.5

Average review count

100

% Hotels top rated

18%

% Hotels bottom rated

3%

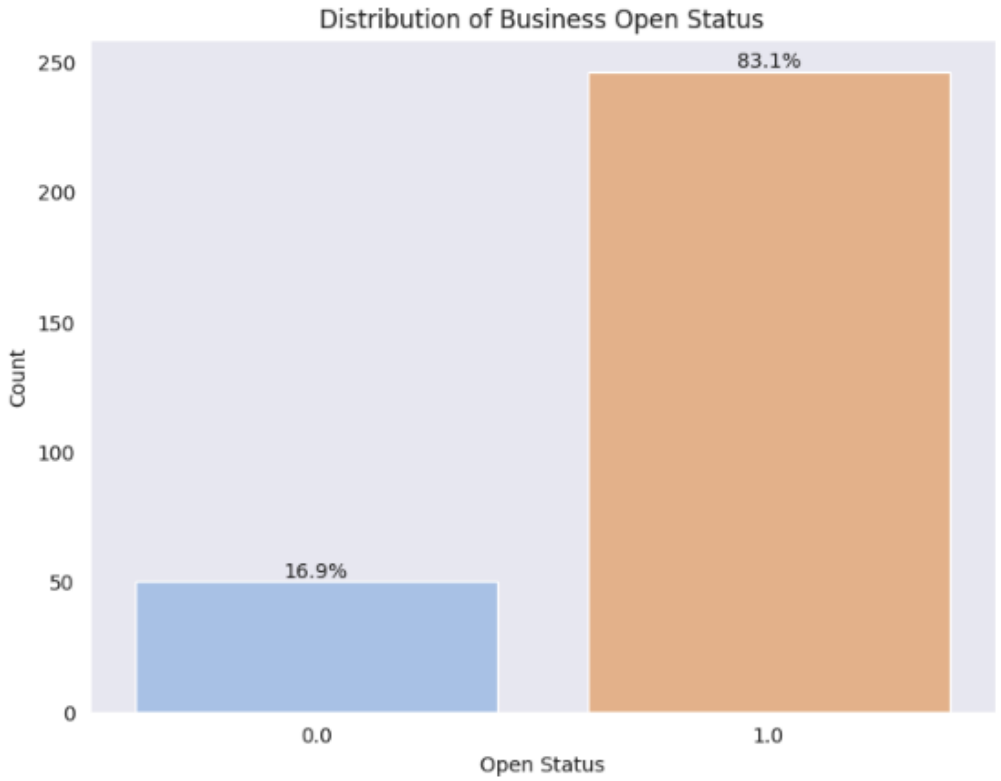
% Hotels top review count

16%

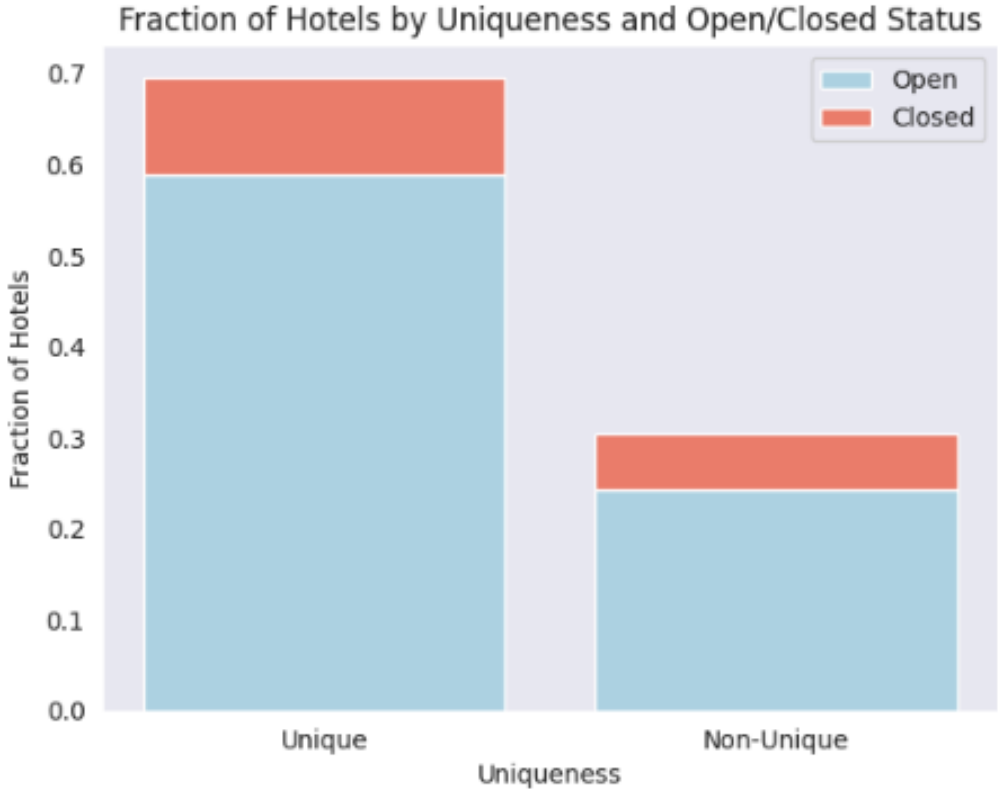
% Hotels bottom review count

1.7%

Distribution of open v/s closed hotels in Louisiana based on if they are part of a hotel chain or not

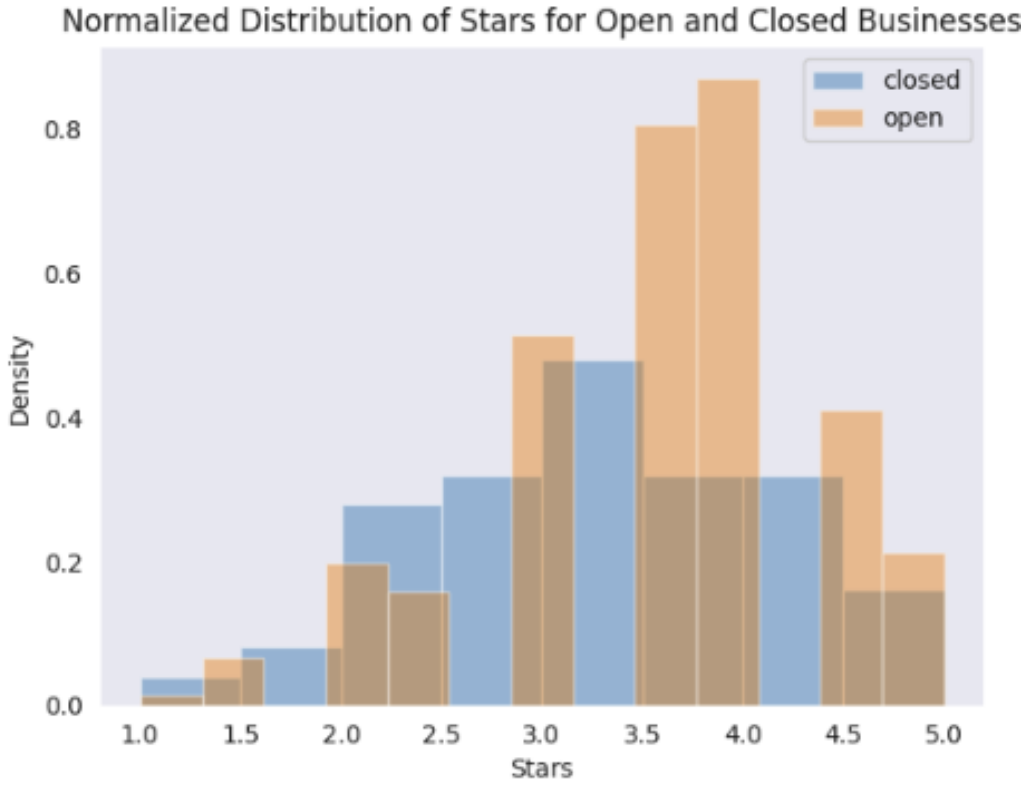


- Imbalance dataset of hotels based on if they are open or not

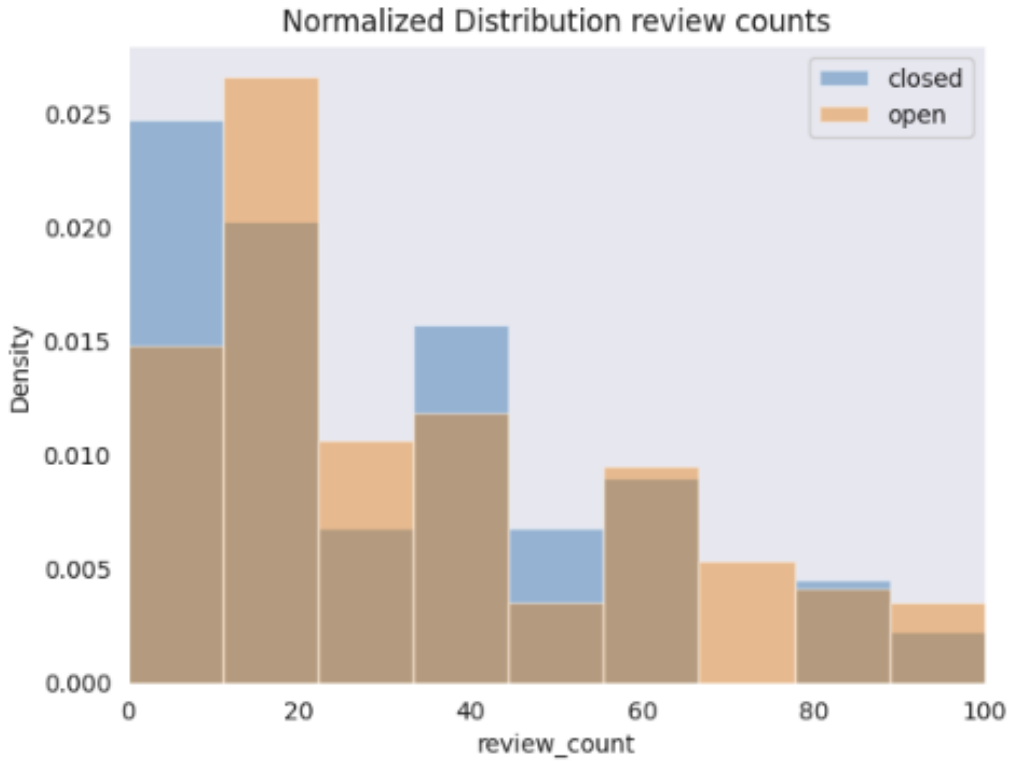


- Unique hotels ie not a big hotel chain compared to a hotel backed by a reputed brand such as Marriott, Hilton etc.
- Not a major differentiator of whether a hotel will close or not

Distribution of open v/s closed hotels in Louisiana based on average star rating & total review count

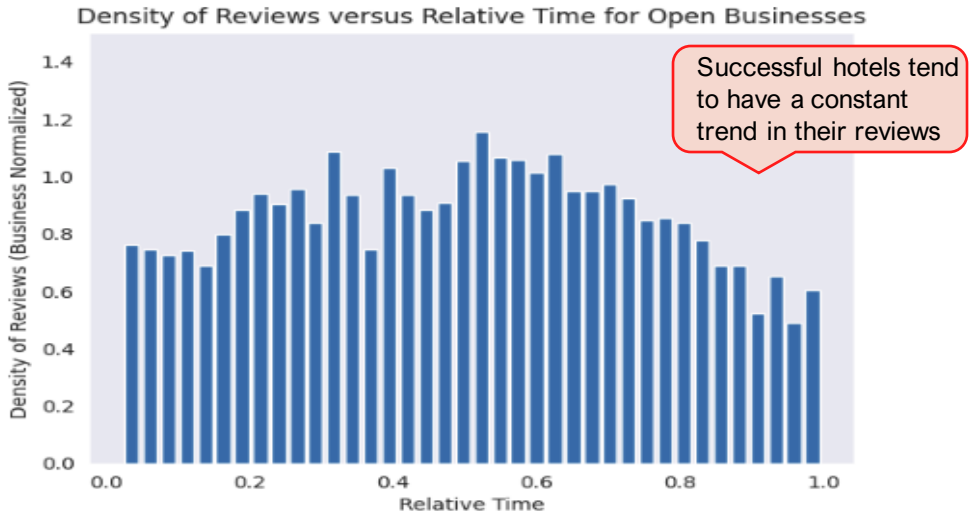


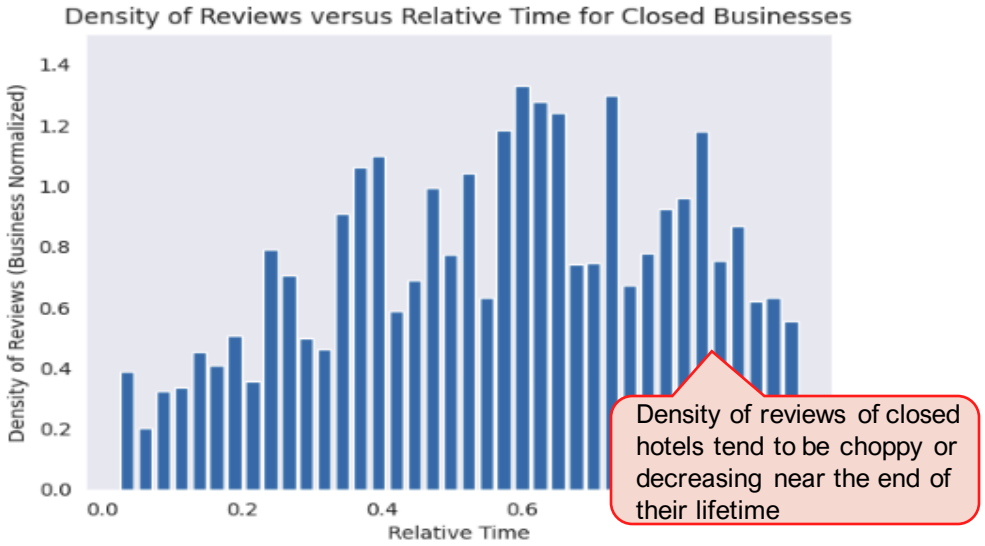
- Average of stars of closed hotels: **3.49**
- Average of stars of open hotels: **3.64**
- Stars alone is not a good predictor of a hotel closing

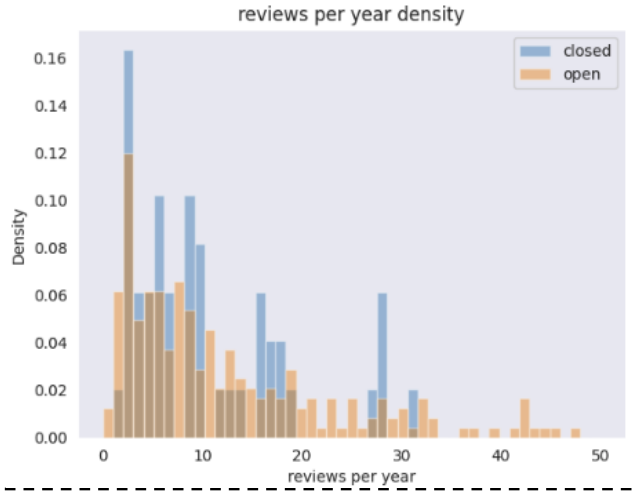


- Review counts of closed hotels are concentrated around a lower number of reviews.

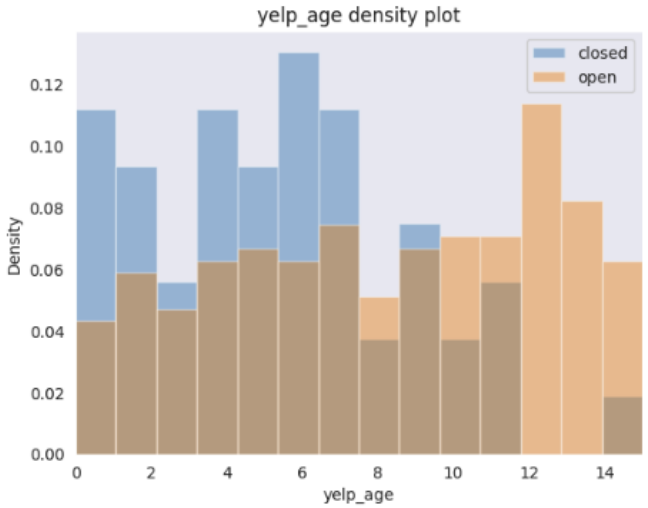
Distribution of open v/s closed hotels in Louisiana based on density or reviews over time, per year and age of hotel







- Reviews per year are higher for hotels which are open.
- Open hotels reviews per year range from 0-50
- Closed hotels have reviews per year from 0-30.



- Older hotels tend towards not shutting down compared to new hotels based on yelp age.
- This suggests that hotels that are open for a long time, and less likely to shut down.

Analytical methods utilized to predict hotel closure in Louisiana

Linear Logistic Regression

- Features
  - Density of reviews
  - Average star rating
  - Relative number of other hotels around the hotel
  - Yelp age
  - Hotel chain or not
- Plot responses, visualize confusion matrix and ROC curve

Sentiment analysis

- Features:
  - Reviews converted into matrix of TF-IDF features
  - Text data then becomes suited of machine learning
- Model
  - TF – IDF tokenizer & SVM
- Evaluation Metrics:
  - Confusion Matrix

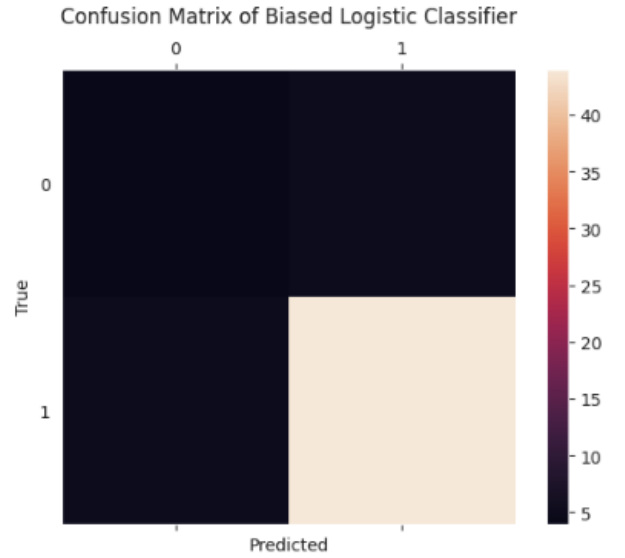
Sentiment analysis methodology using BERT

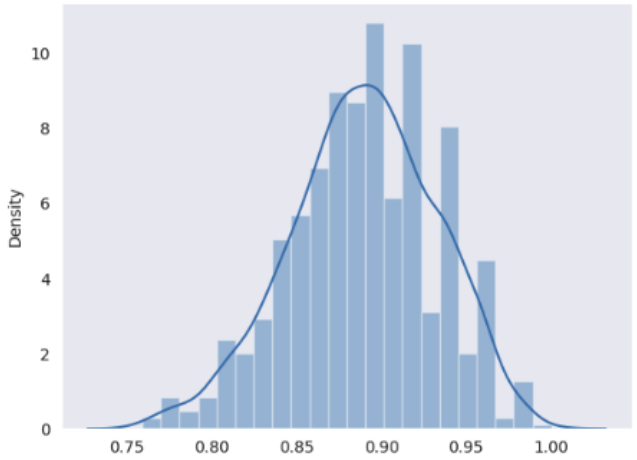
- Features:
  - Scores by group and review sentiment
  - Review Sentiment by Hotel Status
- Model Performance Matrix
  - Confusion Matrix
  - ROC curve and AUC score
- Insights:
  - Ad Study of the effect of sentiment on the status of the hotels.
  - A study of score by review sentiment.

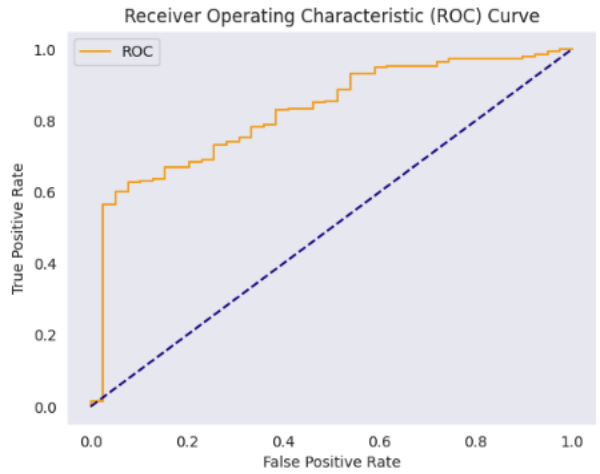
Topic Modeling with LDA

- Features:
  - Bag-of-Words representation of text data.
  - TF-IDF vectorization of text data.
- Evaluation Metrics:
  - Coherence score for topic interpretability.
- Visualization:
  - Word clouds and top words for each topic.
- Insights:
  - Identified key topics/themes in open/closed hotels based on customer reviews.

Linear Logistic Regression







- In confusion matrix, we see hotels which are open have been correctly classified mostly.
- Some level of misclassification by the logistic regression chart. In a perfect world, all blue points should lie above the line and all orange below.
- ROC curve is promising with a low false positive rate
- **Mean precision = .889**

Sentimental analysis based on TF-IDF+ SVM & BERT

	precision	recall	f1-score	support
0	1.00	0.02	0.03	303
1	0.91	1.00	0.95	2938
accuracy			0.91	3241
macro avg	0.95	0.51	0.49	3241
weighted avg	0.92	0.91	0.87	3241

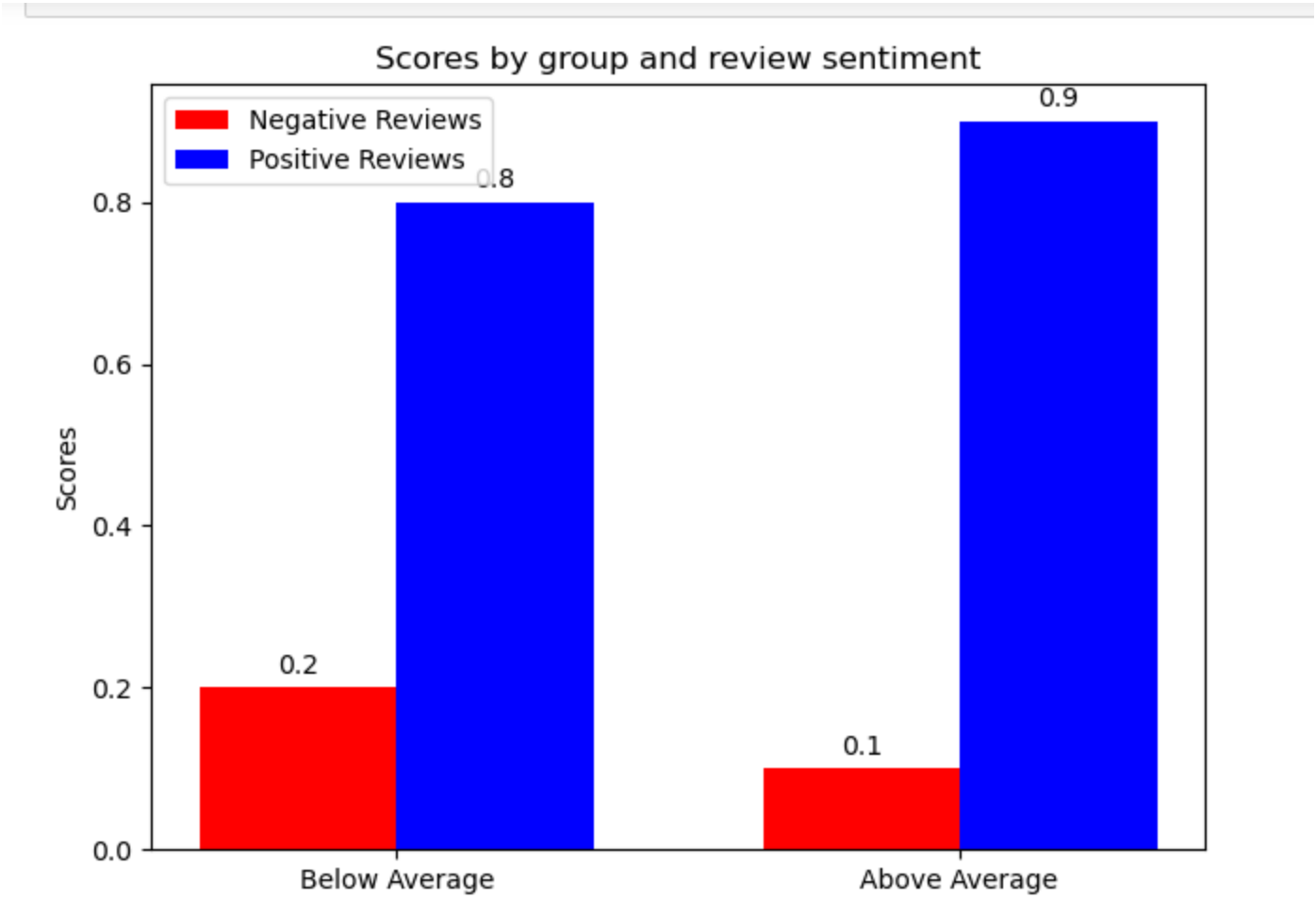
Accuracy of **91%** can correctly predict the open or closed for TF-IDF + SVM

```
- loss: 0.0645 - accuracy: 0.9897
```

We got an accuracy of **98.97%** for our BERT model, and for having a higher accuracy we chose to do sentiment analysis with BERT methodology.

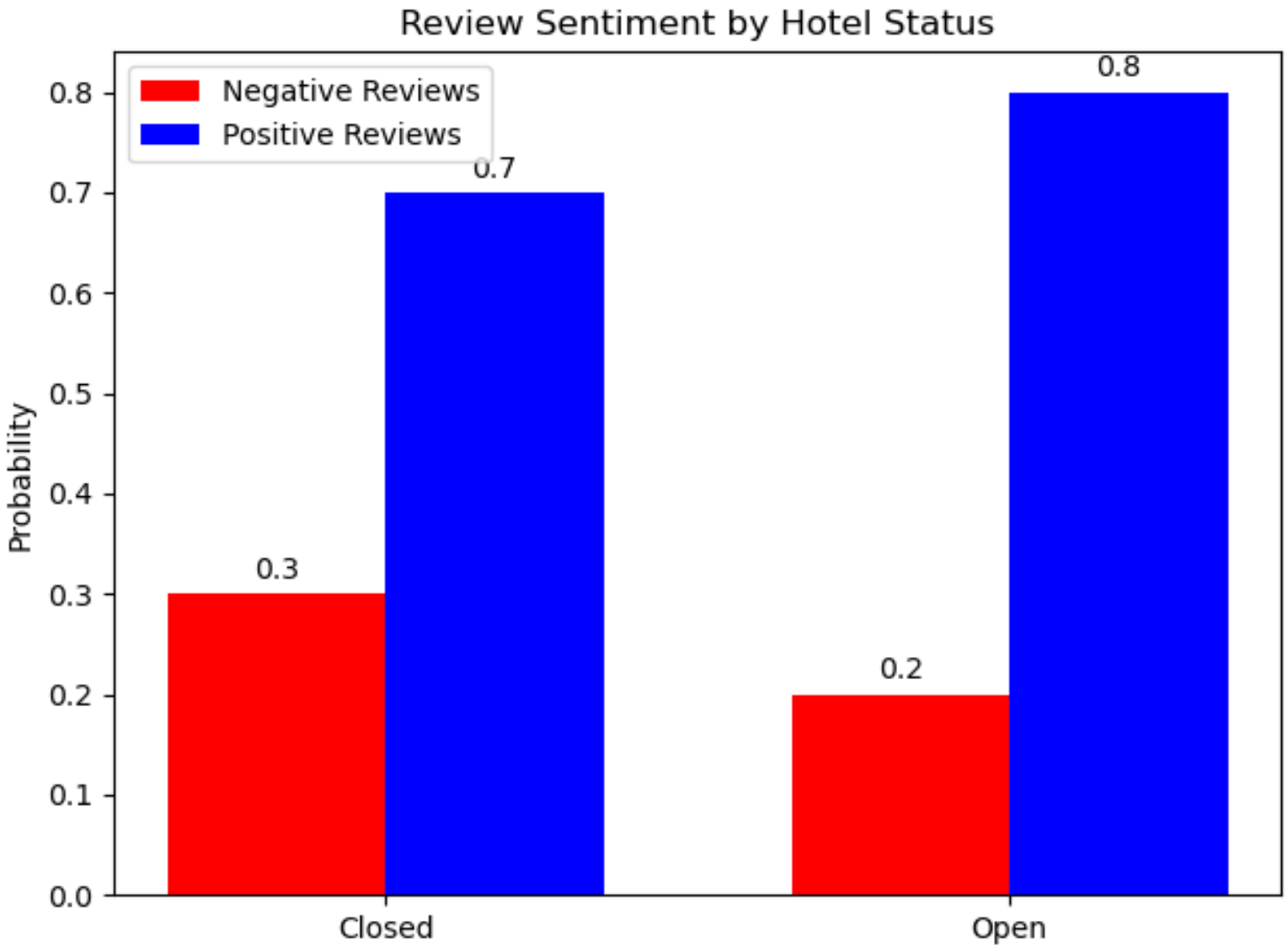
BERT: Review Sentiments Distribution Across Star Rating

- The "Above Average" group shows a significant disparity in review sentiments.
- Negative Reviews receive a notably low score of 0.1 in this group.
- Positive Reviews markedly dominate with a high score of 0.9, highlighting a strong preference or higher satisfaction among reviewers.

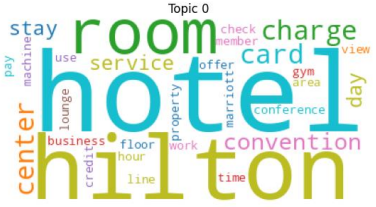


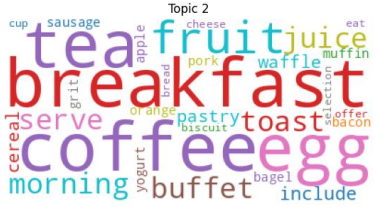
BERT: Review Sentiments Distribution Across Star Rating

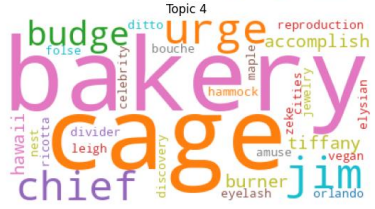
- A higher blue bar in comparison to the red bar for 'Open Hotels' suggests a predominantly positive guest experience.
- Conversely, if the 'Closed Hotels' exhibit taller red bars, this could imply a correlation between negative reviews and hotel closures.
- The comparison offers insights into the potential impact of customer feedback on the operational status of hotels.

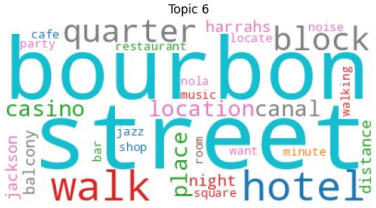


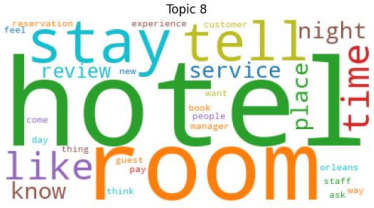
Topic Modeling with LDA for hotels open

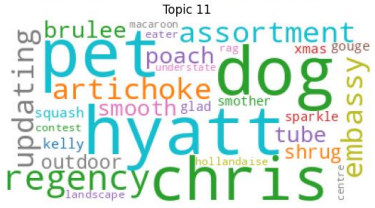


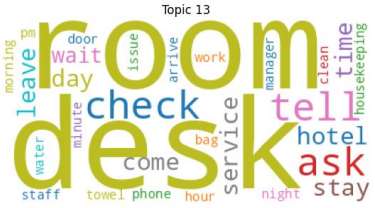


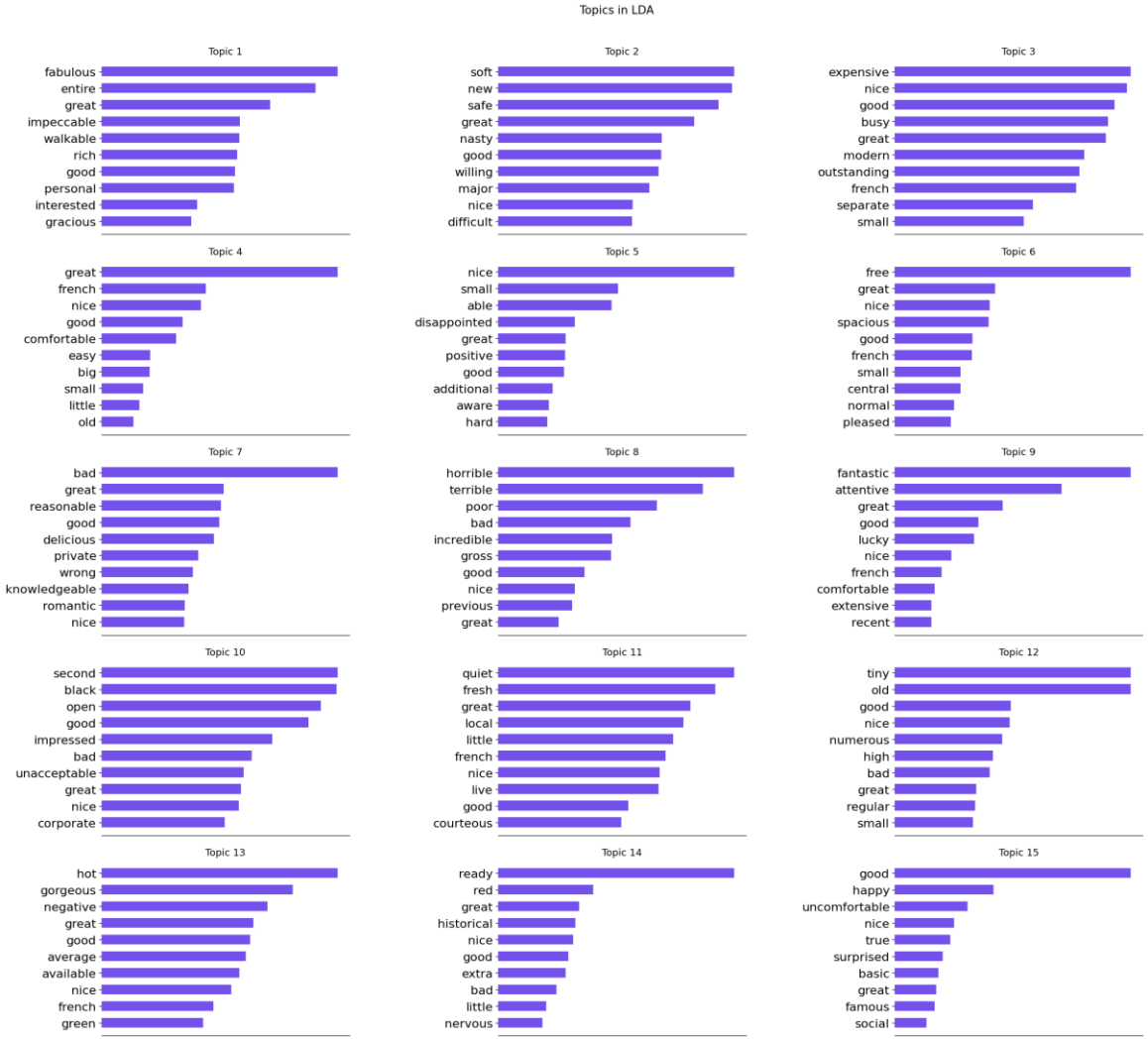






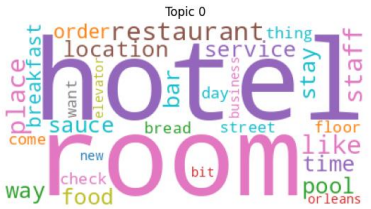


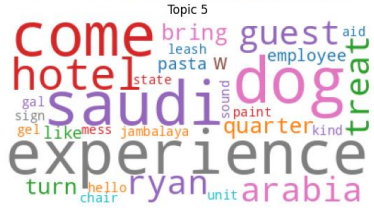


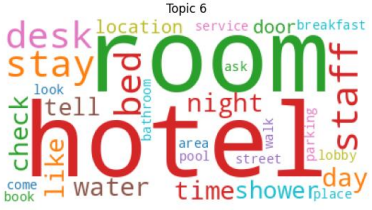


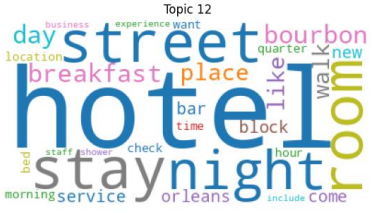
Open Business Themes: Hotel Experience | Food and Dining | Hospitality Services

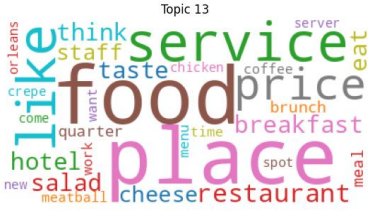
Topic Modeling with LDA for hotels shut down

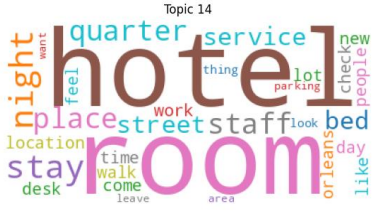


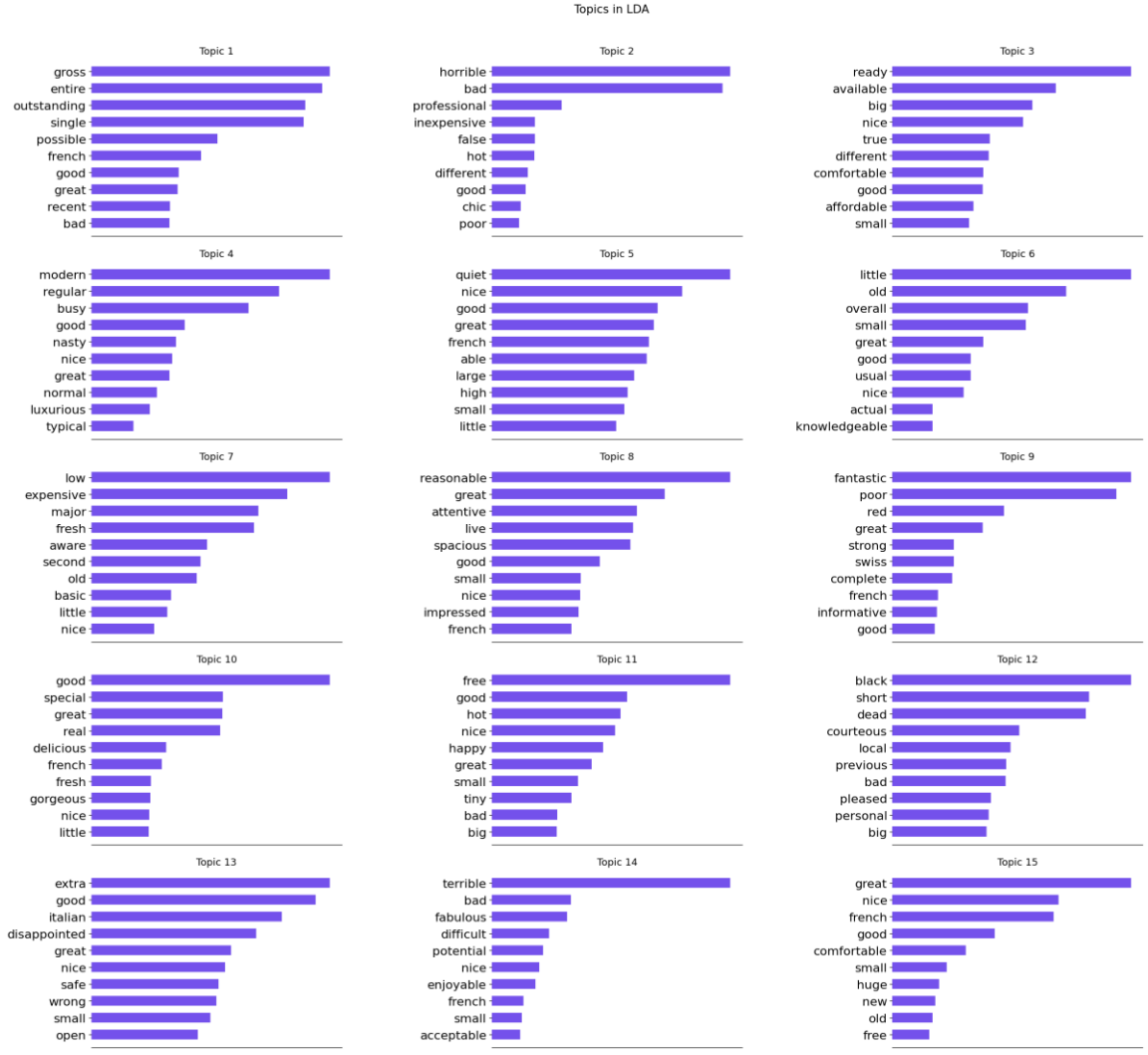












Closed Business Themes: Room and Location Concerns | Dining and Food Service | Customer Experience

Conclusion

Reviews & customer satisfaction

- Recent reviews correlate with a hotel's likelihood of staying open, indicating ongoing interest and potential sustainability.
- Higher star ratings are linked to a business's longevity, reflecting customer satisfaction and continued support.

Age of hotel

- The correlation between the age of a hotel and its success on Yelp is positive.
- Older hotels are likely to have established effective strategies for maintaining profitability, given their track record of success

Hotel location & room concerns

- Focus on improving the location-related aspects such as accessibility, safety, & proximity to attractions
- Prioritize addressing room-related location concerns for best customer experience.

Hotel & dining experience

- Themes identified in LDA recommend to invest in enhancing breakfast and dining options to attract guests and improve overall satisfaction.
- Ensure efficient and friendly service in bars and restaurants to create memorable dining experiences.

Limitation and future considerations

- Survivorship bias may affect analyses of hotel closures in Louisiana, as datasets typically include only recently closed establishments, potentially distorting conclusions

- Discrepancies between recorded ages and actual ages of hotels in Louisiana could impact correlations between age and longevity, especially considering hotels might operate for some time before garnering reviews.

- TF-IDF relies on the Bag-of-Words model, which treats each word independently and ignores the context and semantics of the text. This may lead to loss of important information and nuances present in the reviews, whereas BERT is a more robust model.

- Topic Modeling with LDA – for open hotels:
  - A coherence score of **0.554** indicates a moderate level of topic interpretability.
  - Model score can be improved with different hyperparameters, ensemble methods and fine tuning.

- Topic Modeling with LDA – for closed hotels:
  - A coherence score of **0.313** indicates a lower level of topic interpretability compared to higher scores.
  - Model score can be increased with increase in data quality. Only 1578 datapoints when compared to 14609 datapoints for open hotels.
- BERT modelling is considered a black-box because its internal workings are complex and not directly interpretable to stakeholders

The team

appreciates your time and support

Enabling better decisions

Agile | Accurate | Comprehensive

#LetsTalkProblems