TEXT ANALYTICS IN UK HOTEL REVIEWS

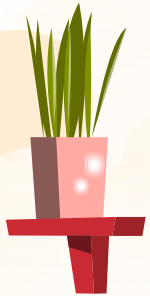
Group Violet

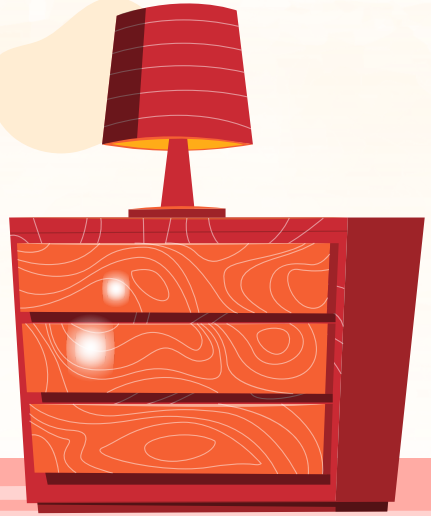
Jinglei Liu Wenjing Xia
Yuqi Zhang Zhewen Liu

TABLE OF CONTENTS

01

BUSINESS PROBLEM





BUSINESS PROBLEM

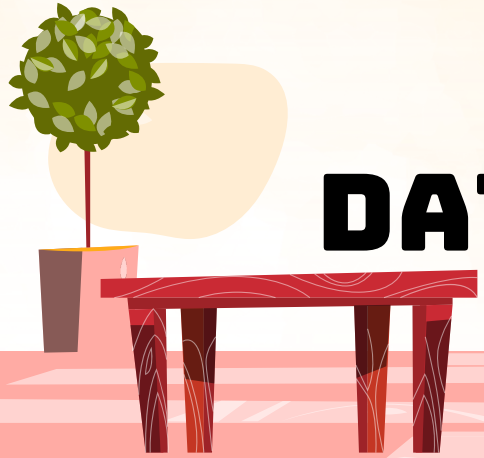
According to the research conducted by SiteMinder:

- 81% of travelers frequently read reviews before booking a hotel
- 91% of travelers (who frequently read reviews) at age 18-34 trust reviews as much as personal recommendation
- 79% of them will read 6-12 reviews before making a decision

Review can be a property utilized by businesses to improve revenue. This project aims to find solutions to improve financial performance for hotels using text analytics techniques.

02

EDA & DATA CLEANING



DATA SOURCE

- We found our dataset from Kaggle
- We selected reviews for UK hotel and within a year from hotel dataset, containing 124,881 rows and 17 columns
- The focus was on the columns of positive reviews, negative reviews and reviewer scores

Source: https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe

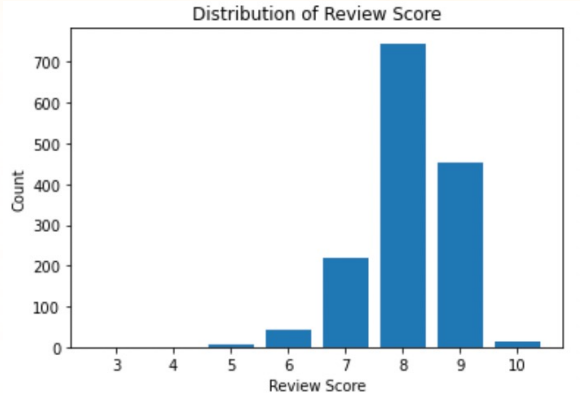
EXPLORATORY DATA ANALYSIS

Trend:

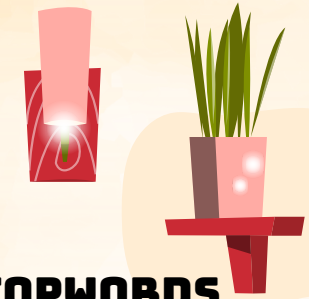
- The review score deceased at the beginning and then increased
- The first quarter in 2017 reached the highest and then decreased
- July in 2017 was the lowest and then increased
- People tend to give high scores during holidays

Distribution:

- Most scores are between 7 to 9



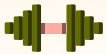
DATA PREPROCESSING



LOWERCASE

Change reviews to lowercase

TEXTACY



Use Textacy to regularize numbers, emojis, currency symbols, and punctuations

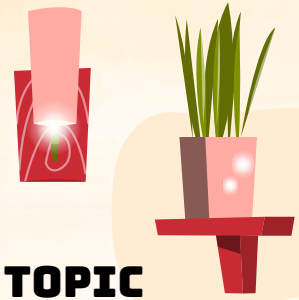
LEMMATIZE

Lemmatize each token and reduce number of features

STOPWORDS

- Remove stopwords from nltk and genism's list
- Add customized stopwords by looking at the 100 most common tokens in reviews
- Keep words like "not", "no", etc. for further sentiment analysis

DATA PREPROCESSING (CONT.)



REGEX

Use RegEx to correct wrong-spelling tokens

N-GRAMS

Draw quick insights into key words by looking at tri-grams

TF-IDF

Use TF-IDF calculations to get important phrases

TOPIC MODELING

Use topic modeling to discover latent topics from corpus

IMPORTANT PHRASES
POSITIVE REVIEWS

FACILITY

Free Tea Coffee
Free mini bar

ROOM

Room modern, tidy, spacious
Bed comfortable, big size

SERVICE

Staff attentive, polite, helpful
welcome, professional
Friendly reception staff

BREAKFAST

Breakfast good selection
Breakfast good value

LOCATION

Royal Albert Hall, view Eiffel Tower
close Tube Station, central
easy access, easy walk distance

IMPORTANT PHRASES
NEGATIVE REVIEWS

FACILITY

Tea & coffee making facility not work well

ROOM

Room is small, unclean
Bed is not comfortable
AC makes much noise

SERVICE

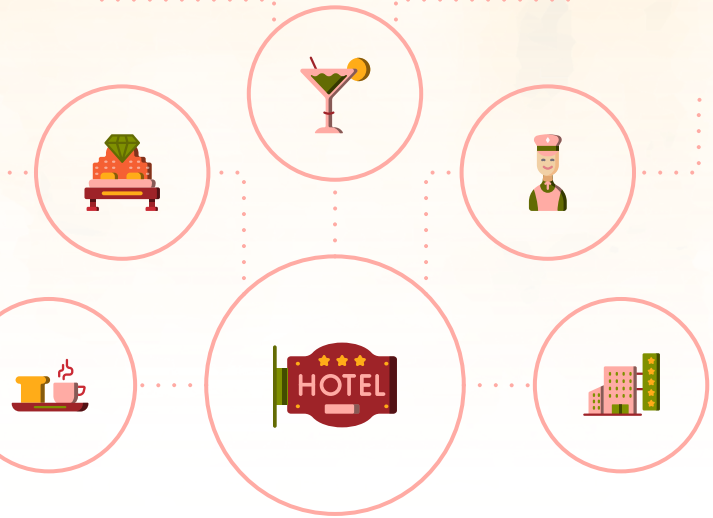
Staff not helpful
Poor room service
Wait time too long

BREAKFAST

Breakfast is not good
Breakfast is too expensive

LOCATION

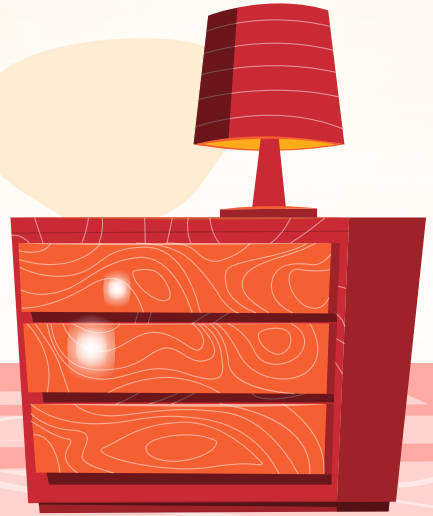
Location is far from city center



IMPORTANT PHRASES

03

MODELING



MODELING RANDOM FOREST

Combine positive and negative regex into a new column “Review”

Data Labelling

- Average Score>=7: Good
- Average Score<7: Poor

Data Encoding

- Good:1
- Poor:0

Train & Test Dataset Split: 80% Training, 20% Testing

Vectorization: TF-IDF

RandomForestClassifier

- Accuracy: 98.48%

MODELING RANDOM FOREST

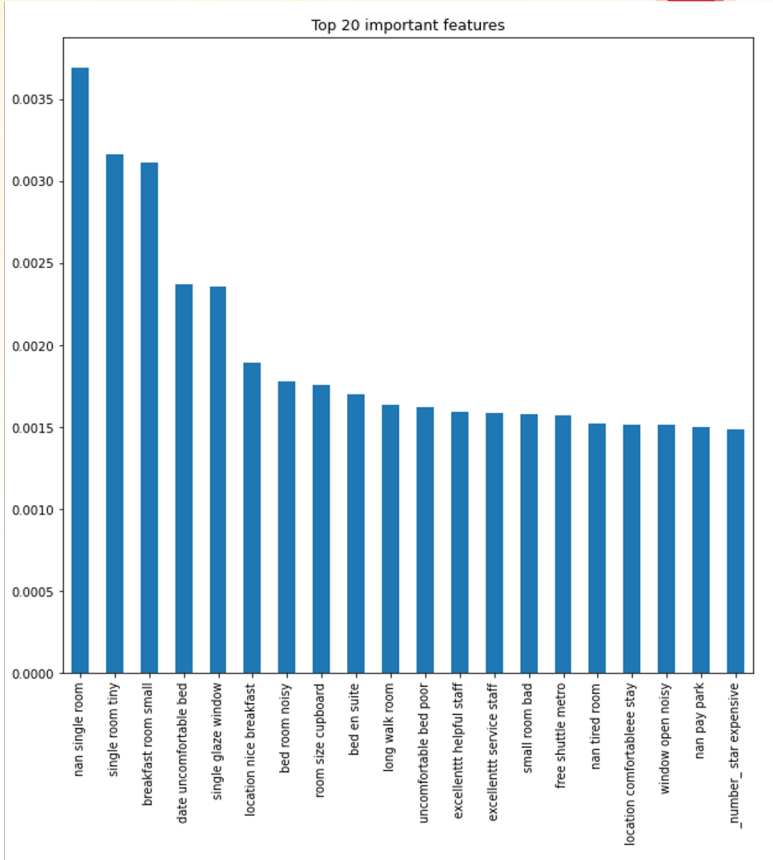
Top 20 Important Features:

POSITIVE

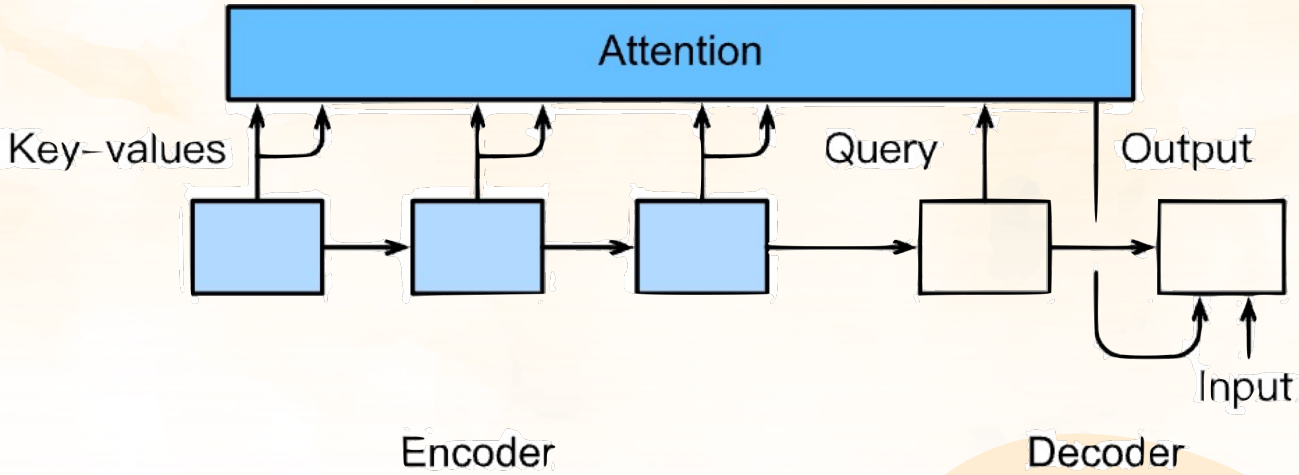
excellent staff
free shuttle metro
location

NEGATIVE

tiny room
noisy
uncomfortable bed
expensive



SENTIMENT ANALYSIS MODEL BASED ON ATTENTION MECHANISM



Attention model can improve LSTM by inserting an attention layer. It extracts the significant aspect of each review and uses them to predict sentiment.

MODEL STEPS & RESULTS

1. Data Processing
2. Load in GloVe Vectors
3. Text Tokenize
4. Load in Embeddings Matrix
5. Define LSTM Model
6. Train LSTM Model
7. Define LSTM Model
8. Train LSTM Model

Result: The accuracy rate increased by 0.48% from 96.68% to 97.16% after using self-attention model.

04

INSIGHTS & RECOMMENDATIONS

INSIGHTS & RECOMMENDATIONS

Guest focus, coming from terms in positive and negative reviews:

- Train reception staff and greet politely with free tea or coffee are good starting points
- Upgrade room for loyal guests
- Change hard or soft mattress and pillow based on guest preferences
- Install free high-speed WIFI, lower the noise from AC and provide a mini bar
- Add more choices for breakfast, offer free or discount to loyal guests
- Provide free shuttle to the airport, tube station, and famous scenic spots

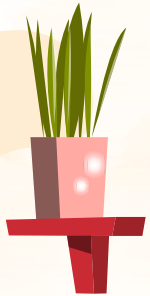
Relate hotel industry trends with the terms:

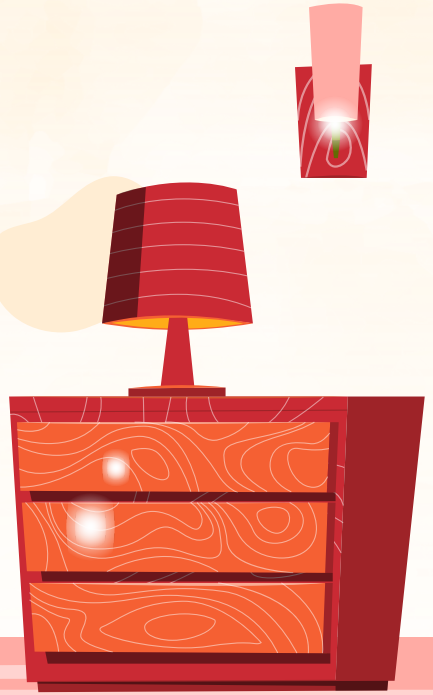
- Wait time too long <- use robot staff to automate check-in and check-out
- Poor room service <- use technology such as artificial intelligence to remember guests' preferences and personalize their room experience
- Modern <- build smart room functions such as adjusting the thermostat, controlling the TV and entertainment systems, or raising the blinds; implement smart controls such as Amazon Alex and Google net to automatically answer guests' questions to improve guest experience and differentiate from competitors

Perform our sentiment analysis model on reviews in social media

05

RETURN ON INVESTMENT





ASSUMPTIONS ABOUT ROI

$7.8B

Agency Revenue for Hotel Booking: Based on the 2017 financial report of Booking Holdings, the total agency revenue was $9.7B and the hotel booking revenue was 80% of agency revenue, which is $7.8B. (source: booking.com)

$52B

Hotel Revenue : Agency Revenue for Hotel Booking / 15% commission rate ≈ $52B. (source: booking.com)

$600K

Investment: We assume the investment for text analytics will be the cost of four NLP engineer plus some software and hardware. Based on Glassdoor, the average salary of NLP engineer is $130K. And the costs of software and hardware are estimated to be $80K.

22%

Percentage of Revenue Lost for Low Score: Based on SiteMinder, 32% of travelers eliminated those with a rating below 7. There are 69% (1019/1483) of hotels in our dataset below 7. 32% * 69% ≈ 22%. We assume that the revenue lost also equals 22% of the total revenue.

0.008%

Increase Rate of Revenue: We assume that, with other things being constant, the increase rate for revenue will be at least 0.008% of total revenue if all hotels can increase their scores above 7 via text analytics

18%

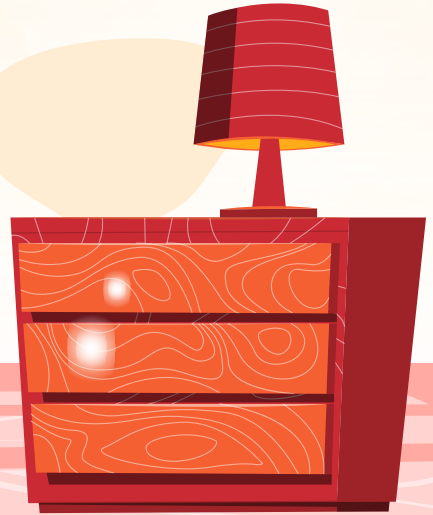
Net Income Rate: Based on the 2017 financial report of Booking Holdings, the rough net income rate was around 18% of the total revenues (2,340,765/9,714,126)

27.46%

Based on the previous assumptions,

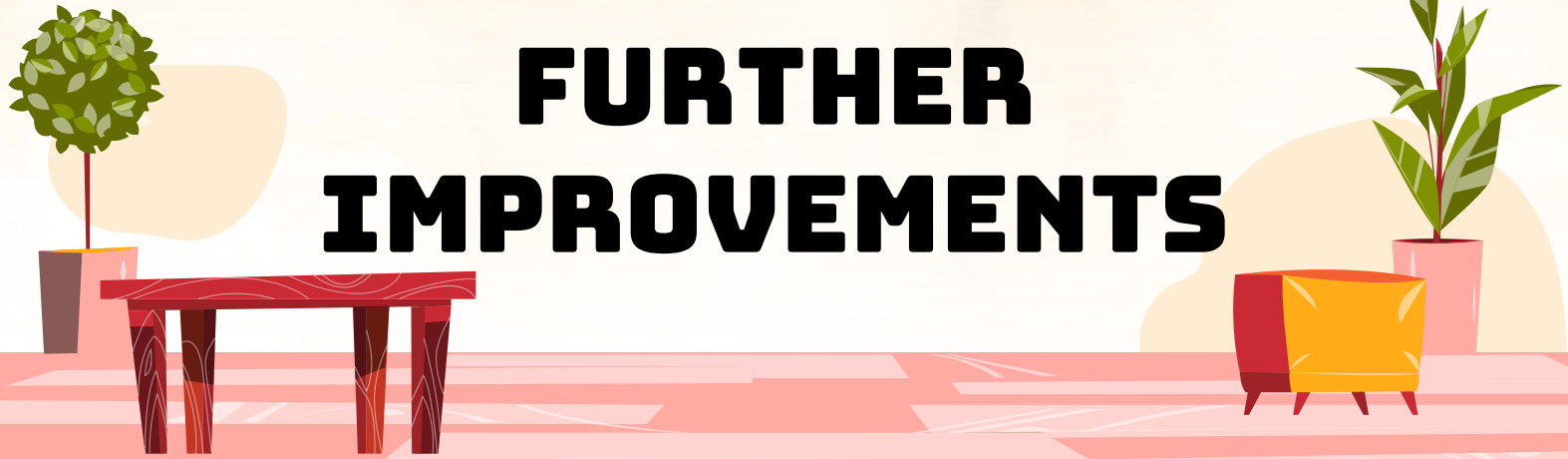
ROI = Net Income / Investment
= (0.008% * 18% * 22% * $52B) / $600K
= 27.46%

**Net Income: Increased Revenue from Text Analytics*



06

FURTHER IMPROVEMENTS



FURTHER IMPROVEMENTS

COMPLEX MODELS

Train more complex models, such as hugging face, multiple layers Neural Networks

OTHER COUNTRIES

Further analyze text reviews in hotel industry in other countries

LEARN FROM OTHERS

Do more research and learn about how successful hotels improve daily management through text analytics

DIFFERENT IMPLICATION

Try different text analytics methods, such as review classification model and genre classification model

THANKS!

REFERENCES

- https://www.kaggle.com/code/jonathanoheix/sentiment-analysis-with-hotel-reviews
- https://www.siteminder.com/r/hotel-trends-hotel-hospitality-industry/
- https://www.siteminder.com/r/hotel-reviews-manage-online-property/
- https://partner.booking.com/en-us/help/guest-reviews/general/everything-you-need-know-about-guest-reviews