Trip Advisor Hotel Reviews: Text Classification Model

Ross Dickinson, Diogo Cobos, Elif Cankaya, Jungwon Zang, Michael Fornito & Hong Liu

I. Background

9 out 10 customers heavily rely on satisfaction rate and reactions of previous customers before making a purchase decision (Oberlo, 2021). Given the profound influence previous customers product or service satisfaction reviews, it is imperative for businesses to assess and evaluate the social media reviews on their services or products.

Study Purpose

The aim of this project is to introduce deep learning, Recurrent Neural Networks (RNN), as a tool for analyzing and evaluating customer reviews in the hotel industry.

II. Methodology

Hotel reviews and reactions gathered from a popular travel accommodation website, TripAdvisor.

- There are 20,491 reviews.
- Each text review is linked to a rating between 1 through 5 (the highest rating).

Recurrent Neural Networks (RNN)

- Usually used for sequential data
- The layers give the model a short-term model
  - ✓ Allows to predict the next model
  - ✓ The past data kept isn't fixed but depended on weights
- RNN is used in sentiment analysis, sequence labeling, speech tagging, etc.
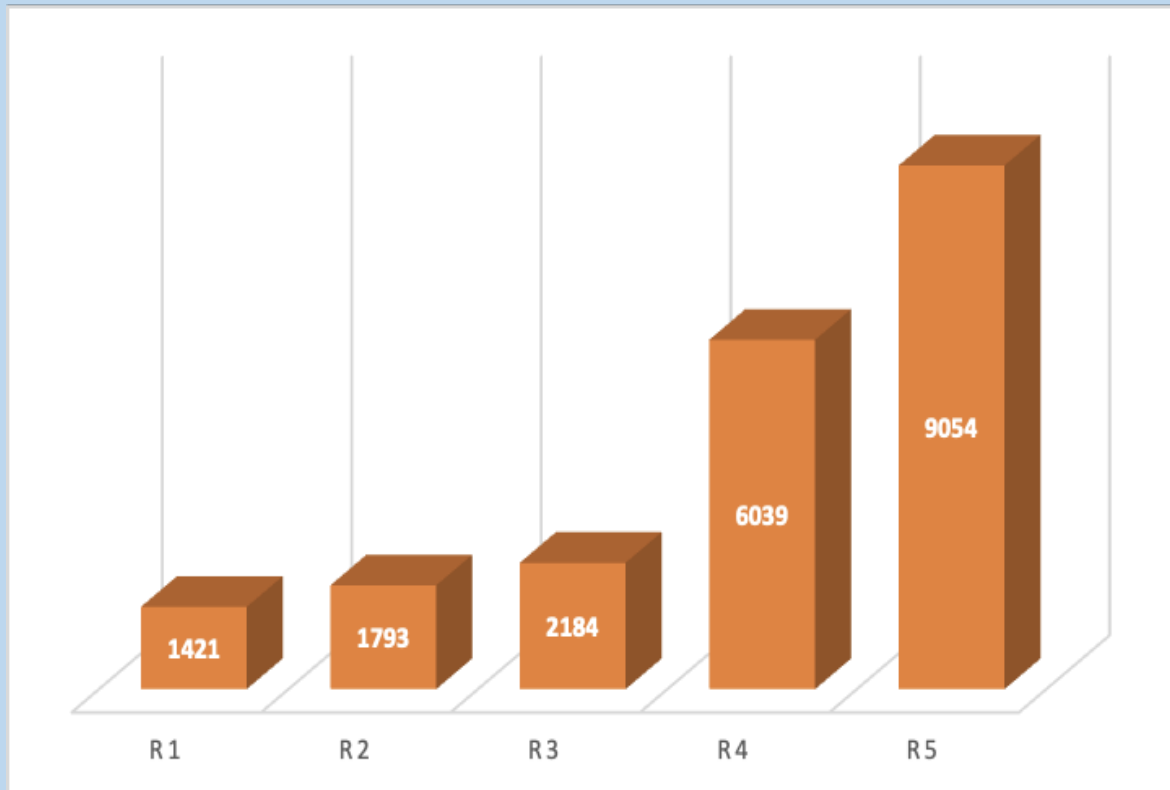
Data Preprocessing

Raw Data   Data Cleaning (Removing stop words stemming)   Data Processing (Tokenize, Convert to Sequence & Pad)

Exploratory Analysis

The bar charts reflecting the frequency of top used words in the reviews based on each rating.

The distribution of review is negatively skewed, and nearly half of the reviews corresponds to the highest rating. 74% of the reviews are positive (ratings 4 & 5).

As reflected 'hotel' and 'room' are the most frequently used two words for each rating.

Frequency of Words based on Each Rating

Frequency of Words
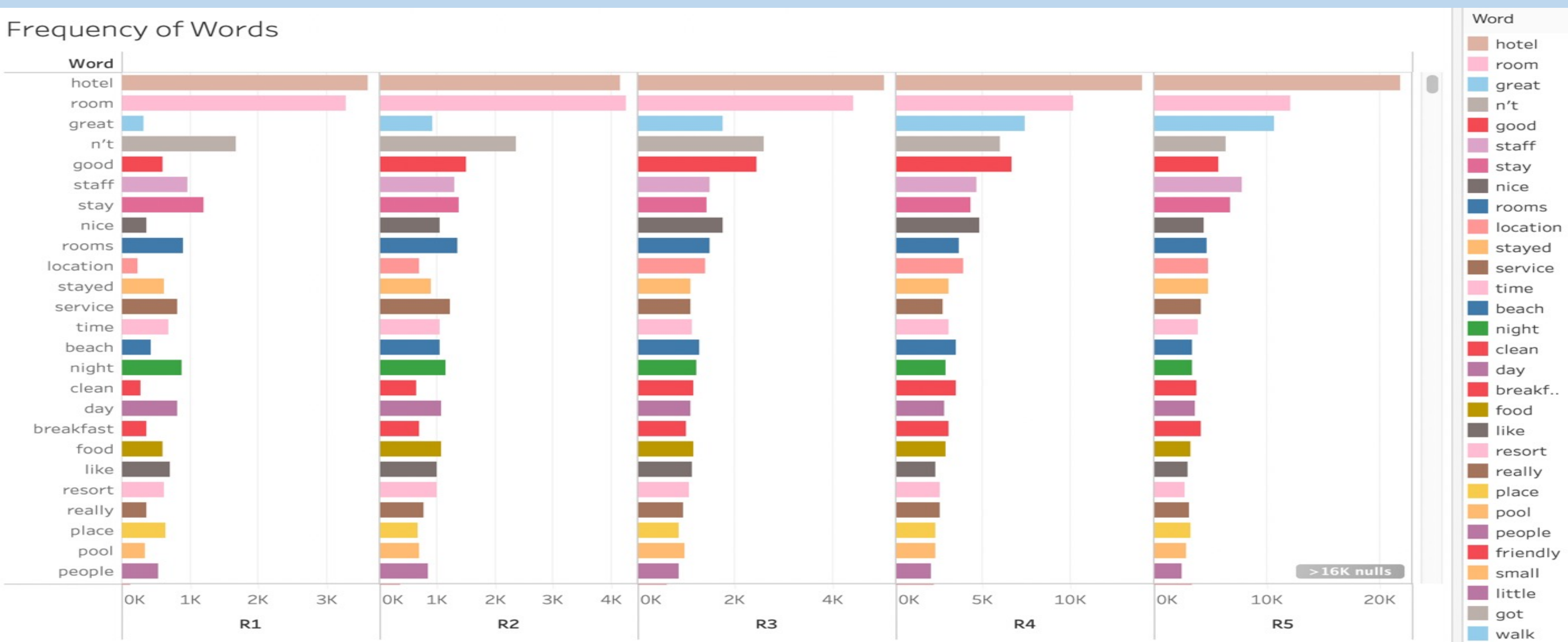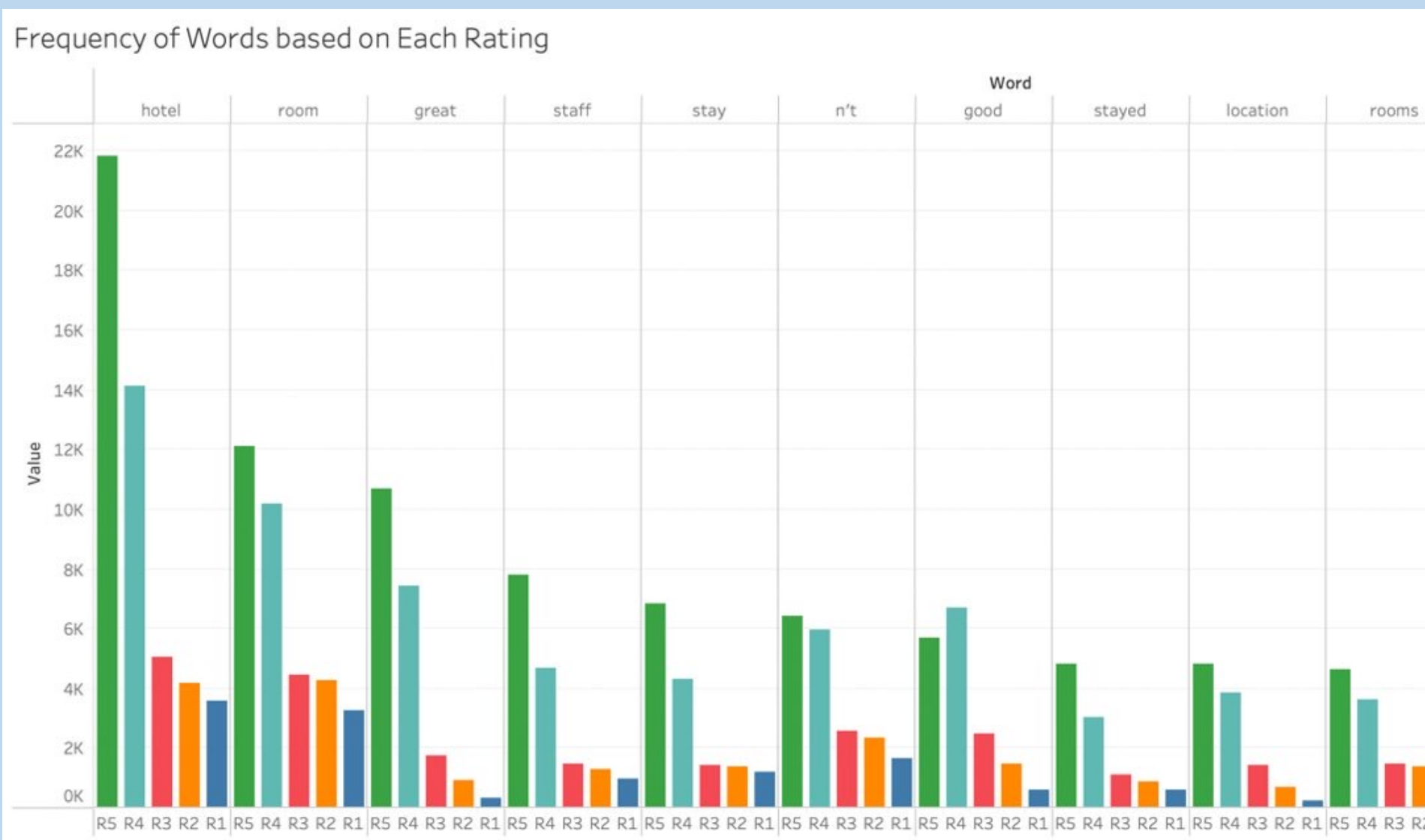
III. Result

A binary, a 3-label and a 5-label models were run to conduct sentiment analysis of hotel reviews using mainly RNN Long-Short Term Memory (LSTM) layers, which can be implemented through the Tensorflow library.

Tuning	Binary		3-Label		5-Label	
	Train	Val	Train	Val	Train	Val
NO	1.0%	0.89%	81.16%	82.5%	55.86%	60.37%
YES	-	-	84.45%	86.99%	59.95%	63.15%

5-Label RNN

The model contains five possible hotel ratings range from one star, very bad review, to five stars, excellent review.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 400, 128)	6400000
dense (Dense)	(None, 400, 128)	16512
bidirectional (Bidirectional)	(None, 400, 128)	98816
bidirectional_1 (Bidirectional)	(None, 64)	41216
flatten (Flatten)	(None, 64)	0
dense_1 (Dense)	(None, 128)	8320
dense_2 (Dense)	(None, 5)	645

Total params: 6,565,509
Trainable params: 6,565,509
Non-trainable params: 0

Hyperparameter tuning was conducted and the entire training and validation process took about 36 hours using GPU settings on Google Colab Pro.

IV. Conclusion & Future Work

- Accuracy: Binary > 3-Label > 5-Label
- Validation accuracy for 5-Label much harder to improve
- Observe the effect of different vocabulary length as well as embedding dimensions
- Tuning for optimal batch size