European Luxury Hotel Sentiment Analysis Using Naive Bayes and Deep Neural Networks

Chih-You Chen & William Cheung
Fordham University, Bronx, NY 10458, USA
cchen201, wcheung5@fordham.edu

***Abstract*— Travel is becoming more acceptable to the general public as technology grows. Given the nature of traveling to an unfamiliar place, information such as reviews has become critical for many travelers and in turn hotel businesses. With the increased web traffic of reviews, it's important to be able to translate the reviews to the sentiment that the reviewer has to the establishment. In order to achieve this, we implemented convolutional and GRU neural network trained on reviews and also other user features in order to rate the person's feelings (good, neutral, or bad). This can be useful in situations such as Twitter where there are text comments but no explicit score or rating is given. With GRU, we were able to archive a 0.63 accuracy on the three-class classification problem while non deep learning baseline model Naive Bayes was only able to achieve 0.41 accuracy.**

I. Introduction

A. Motivation

Travel has evolved into an experience that is increasingly being available to the public. Initially traveling for vacations was only reserved for the super-wealthy because of the financial and time cost to go to different locations. By the 1830s steam trains has enabled more people to travel [1]. This made short distance traveling quicker but vacationing to other countries was still inaccessible to many until aviation. It was only since the 1950s that aviation travel was cheap enough for the general public [2]. A recent study from national geographic projects that air travel will double from 2016 to 2035 [3].

Hotels that housed travelers started in Asia in the 1200s and the idea spared hotels in Europe by the 1500s. These were mainly locally run enterprises and traffic was not frequent. Alone with the travel boom in the 1950s, the hotel industry grew exponentially [4]. Just form 1995 to 2010 the hotel industry grew 61% [5]. Even with the setback from Covid-19, it is projected that in 2021 there will be a recovery of 57.3% [6].

With the boom of the hotel industry, there are so many hotels to choose from. As with the nature of the business, travelers are typically not familiar with the area they are traveling to and need to know which hotel is for them. Reviews and recommendations help travelers but have a huge impact on hotels themselves. With the internet boom around the 2000s, it gave opportunities for sites to develop. From 2010 reviews have blossomed by 20% per year until recently [7]. However, there is a growing tight network of review sites that control a majority of the reviews. Of current travelers 99.7% check reviews before booking vacations [8]. Of those reviews, 78% of them come from Booking.com, TripAdvisor, Google, and Hotels.com [9]. This gives these sites great power over the hotel industry.

B. Contributions

We present two visions which manifested into the text review model and the text review plus background features model as explained in Section II-E. The text review only model was able to achieve a 0.62 accuracy score on the sentiment of reviews classified as one of good, neutral, and bad while background features model was able to achieve similar results at 0.63 accuracy, but its space and information cost made the text review model the better option. The trained model is designed so that the model can be then used on any text-based review to analyze European luxury hotels.

II. Approach

We first need to discuss the setup of the experiments before diving into the performances. After gathering the data we must clean it and then compute features before the learners can start training and predicting data points. Finally, we will also go into the different models we are interested in exploring to develop useful predictions.

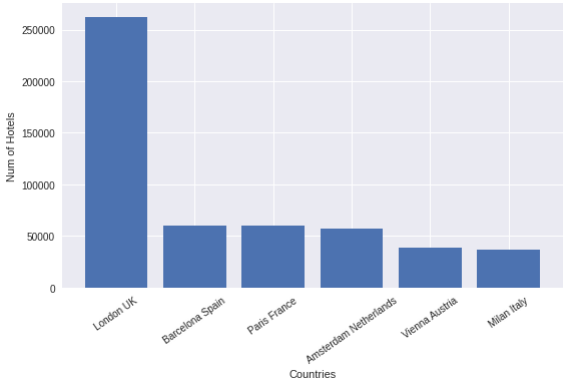

Fig. 1: Demographics of the hotels

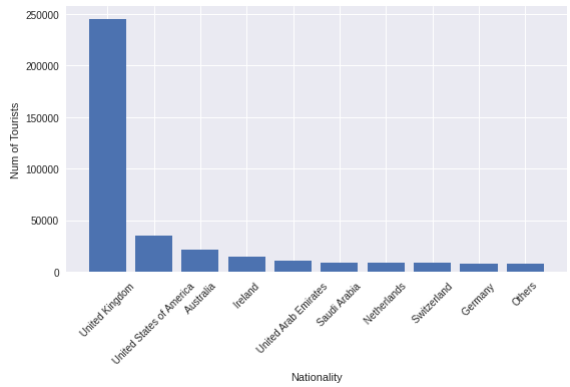

Fig. 2: Distribution of Tourist Nationality

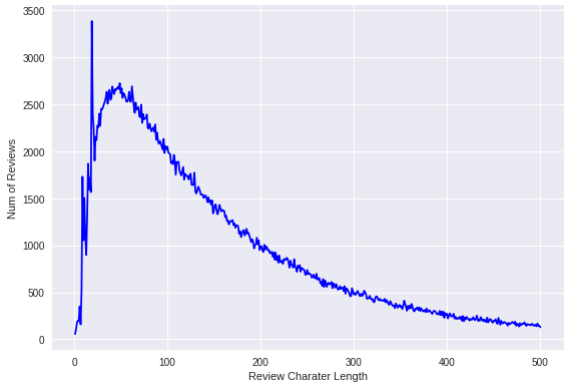

Fig. 4: Distribution of Review Lengths

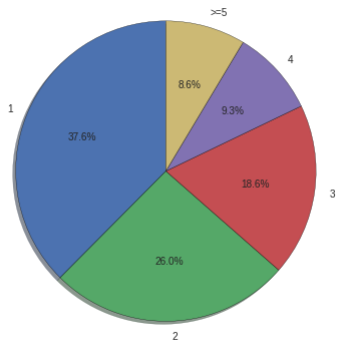

Fig. 3: Break Down of Stay Lengths

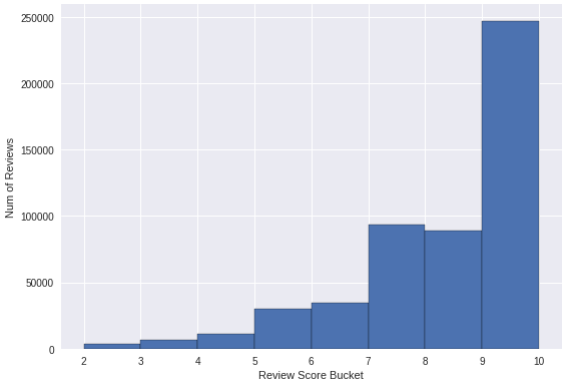

Fig. 5: Distribution of Review Scores

A. *Datasets*

Booking.com dataset: The data set comes from Kaggle [10] and consists of information on 515,000 reviews. The data comes with the following information: hotel address, an additional number of scoring, review date, the average score for that hotel, hotel name, reviewer nationality, negative review, negative review word count, the total number of reviews for that hotel, positive review, positive review word count, number of total reviews this reviewer has given, review score, tags, days since that review, and the coordinate location of the hotel.

Additionally, there were 1,493 total hotels in consideration. The data is specifically on European luxury hotels which defined by booking.com is a five-star property [11].

B. *Data Distribution*

Before diving into the CNN we first look at the distribution of the data to get a better understanding of the data. Firstly, as seen in Figure 1, we can see that the hotels within the dataset come from six locations. All of which are major cities except for the United Kingdom (UK) which is from hotels across the country. Over half of the reviews are on 399 London hotels.

Figure 2 shows a similar picture when looking at the nationality of the travelers. Almost half of the travelers are from the UK. So overall this data set has heavily influenced by UK hotels and travelers.

We also we compositional makeup of the length of stay among travelers in Figure 3. Surprisingly the majority portion of the dataset is of short stays, which we defined as three days or less. This is interesting as by law EU countries have four weeks paid vacation but may decide to take many short vacations [12]. Another factor is there are lots of business trips causing many short stays.

When observing the reviews themselves we noticed in total (negative and positive components combined) a majority are less than 200 characters. The distribution in Figure 4, show that there is a peak at about 50 characters and drops off. This indicates most reviews are a paragraph or less.

In terms of reviews shown in Figure 5, there is a disproportionate amount of 8+ review. This is expected when the hotels considered are deemed luxury five-star establishments. We normalize this to a sentiment analysis described in Section II-E and II-D.

When taking a close look at the reviews, we can get a glimpse of things that contribute to good reviews and those of bad reviews. The word clouds shown in Figures 6 and 7 show the most frequent words used in positive and negative comments respectively. Words

Fig. 6: Word Cloud of Most Common Words in Positive Comments

Fig. 7: Word Cloud of Most Common Words in Positive Comments

that generally trend toward more positive reviews talk a lot about the staff and location. Negative comments generally focus on room size, bathroom conditions, and breakfast.

C. Data Pre-Processing

The columns we utilized includes positive and negative review text , nationality, stays of days and reviewer's location and finally scores as the target values. To consider both positive and negative reviews from reviewer, we first combine the negative and positive comments of each person's stay as a net review. Then, to clean the text data, we first tokenize it, remove stop words and digit tokens and also implement the stemming.

Next, we compute and select influential features before constructing authentication models.

D. Feature Computation

We compute the following sets of candidate features.

- BOW features: We take the bag of words (BOW) approach taking the counts of the vocabulary of the presented dataset using the Sci-kit learn package's CountVectorizer.
- TFIDF features: We take the term frequency–inverse document frequency (TFIDF) using the Sci-kit learn package's TfidfVectorizer.
- Background features : Background features consist of length of stay in days, nationality in country, and location of hotel in city.
- Sequential embedding features: embedding trained by the padded sequential sentences that has the same length using the Keras package's Embedding function.
- Sentiment label: As described previously in Section II-E, there are three division to create three classes of good, neutral, and bad. When reviews are higher than 7.5, it is classified as good, when it is lower than 5, it's classified as bad, and for all the others, it is classified as neutral.

E. Methods

When looking at the data sets we decide that we want to normalize the reviews so that it reflects the performance of luxury hotels against its peers. To achieve this we create three buckets: positive, neutral, and negative based on even distribution of the review. This means we down sample positive and neutral review so that they have the same number of samples as the negative one for training.

Now an in terms of the CNN model we compare two structures. The first is to train CNN on all the reviews. The second is developing works on the assumption that we have access to additional information. So along with text the second type also has reviewer's nationality, length of their stay, as well as the hotel's location. Each have been train with BOW and TF-IDF version of review translations.

For the sequence model, we choose to implement GRU. Same as the experiment we've done in CNN, here we also compare training based on only reviews' text data and training along with others reviewer's features(nationality, length of their stay,and hotel's location) to see if the performance would improve. To generate the embedding for the sequential model, we trained keras' embedding layer and then add a GRU layer based on those trained embedding. And in the last layer, it's a dense layer with softmax as activation function that we used to predict the sentiment class.

To better compare different models, we use Naive Bayes based on both Bag Of Words and TFIDF as our baseline model .

In total we have developed 8 different models.

III. Evaluation

Now that the set up of the experiments have been presented, we can implore on the performances of the models. But first, we have to discuss the methods of training and testing.

A. Training-Testing Set

We tried 5 fold cross-validation where each iteration is four splits in training and one split in testing creating a the 80%-20% split in order to better evaluate the result and prevent the result from biasing toward the selected set of validation and testing data.

B. Performance Measures

We consider the following measures to evaluate the performance of different modeling approaches:

Accuracy (ACC), which is the fraction of predictions that are correct, i.e.,

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Now considering the performance of the combination of weighted accuracy we take the sum of short stay and long stay model times the percentage of the data used to train and test the model for each respective model.

C. Hyper-Parameter Optimization

We first use Sci-kit Learn grid search wrapper to find the optimal batch size, number of epochs, learning rate, and optimizer. Then, we try different numbers of hidden layers and different numbers of nodes in those hidden layers. After finding the optimal models, we perform our last step of the grid search by deciding the optimal activation layer, pooling method, and regularizer. Below is the list of hyper-parameters with their ranges of values that we tune for each model.

- Optimizer: adam
- Batch size: 50,100,256
- Epoch: 20,50,100
- regularizer
  - l1
  - l2
- Node, hidden layers: (128,64)
- Learning Rate: 0.01

Table I shows the best hyper-parameter configuration for each feature set.

TABLE I: Performance of CNN models. The hyper-parameters that were consistent optimal choices were the l2 regularizer, 0.01 learning rate, and adam optimizer. The + indicates that the background features were also used. V ACC is the validation accuracy and T ACC is the Test Accuracy

Model	batch size size	epochs	feature count	V Acc	T ACC
NB BOW+	all(66726)	None	17,106	0.69	0.41
NB TF-IDF	all(66726)	None	17,106	0.70	0.41
GRU	256	20	17,106	0.81	0.62
GRU+	256	20	17,109	0.77	0.63
CNN BOW	100	50	81,265	0.58	0.57
CNN BOW+	100	50	81,268	0.62	0.60
CNN TF-IDF	100	100	37,497	0.66	0.62
CNN TF-IDF+	50	50	37,500	0.68	0.63

D. Model Comparison

In these sections, we will compare the different approaches, the text reviews BOW model and the text review BOW plus background features models. As seen in Table I, the CNN added features model outperforms the previous model by 5.3% from 0.57 to 0.60. With a guessing rate of one third in this three-class classification model. The models show promise in that it achieves 80% better than guessing performance at 0.60. There is a reasonable correlation between hotel sentiment and the text alone in a CNN model.

Among the CNN models TF-IDF performs slightly better than BOW and background features also have positive impact.CNN TF-IDF performs at 0.62 without background features and 0.63 with them. In addition to better performance than BOW, the TF-IDF models saves a lot of space as the vocabulary is 26,666 is lower than the two components of the combined model.

From the evaluation result, we can see that sequential model GRU has the highest accuracy (81%) in terms of training. And in testing process, GRU with background features added and CNN by tfidf has higher accuracy (63%) than others. For CNN, using TFIDF is better than BOW as features. Generally, adding reviewers' background information has slightly better accuracy than not adding, however, the increase of accuracy is all within 3% which is not so significant. Compared with the baseline Naive Bayes model, the deep learning methods all have better testing accuracy, showing that our deep learning methods is better for this problem in our project.

IV. Discussion and Future Work

In this project, we found a possible methods to predict reviewers' sentiment in hotels' review data. This has board applications also on other text-based data like Twitter posts, blogs, and other review sights. Although we find several reviewers' profile(length of stay, nationality, etc.) doesn't significantly help in predicting sentiment, others' information such as reviewers' age, visited time etc. might be useful and can be added

in the future. Plus, several possible adjustment can be considered in future work 1. Increase more various data in the training process 2. Try to remove less stop words to keep more sequential information 3. Try more hyperparameters tuning

V. Conclusions

This work focuses on developing a text-based sentiment analysis of European luxury hotels. The data coming from Booking.com which has the largest hotel review the model design was to make the resulting CNN usable from other text sources such as Twitter and blogs on the same type of hotels. With our CNN and GRU model, we were able to achieve an testing accuracy of 0.63, while the Naive Bayes with TF-IDF performed the best with 0.41 in a three (good, neutral, bad) class problem.

References

[1] "History of the vacation," Accessed: December 2020. [Online]. Available: https://rb.gy/qxsgdf

[2] "What it was really like to fly during the golden age of travel," Accessed: December 2020. [Online]. Available: https://rb.gy/qxsgdf

[3] "As billions more fly, here's how aviation could evolve," Accessed: December 2020. [Online]. Available: https://rb.gy/wzmzwn

[4] "Hotels — a brief history - by jacques levy-bonvin," Accessed: December 2020. [Online]. Available: https://rb.gy/3rcy1o

[5] "What could the next 40 years hold for global tourism?" Accessed: December 2020. [Online]. Available: https://rb.gy/j2zqe2

[6] "How the pandemic is changing 2020 hotel forecasts," Accessed: December 2020. [Online]. Available: https://rb.gy/jllzg1

[7] "Online reviews still important but craze is slowing down," Accessed: December 2020. [Online]. Available: https://rb.gy/s2zdyu

[8] "Report: 78% of all online hotel reviews come from the top four sites," Accessed: December 2020. [Online]. Available: https://rb.gy/uwve0d

[9] "The impact of online reviews on the hospitality industry [infographic]," Accessed: December 2020. [Online]. Available: https://rb.gy/r8euob

[10] "515k hotel reviews data in europe," Accessed: December 2020. [Online]. Available: https://rb.gy/y42b3o

[11] "Luxury hotels in new york," Accessed: December 2020. [Online]. Available: https://bit.ly/2K4RnFk

[12] "On holiday: Countries with the most vacation days," Accessed: December 2020. [Online]. Available: https://rb.gy/mgsxsx