NLP and Machine Learning Modelling for Las Vegas Hotel Reviews

BIA660-Project Report

Team 6

Yuiang Zhai

Hantao Ren

Nitin Gullah

Abstract

In this project, we aim to predict rating of the reviews using sentiment analysis, topic modeling and machine learning techniques on data from booking.com, a travel fare aggregator website and travel metasearch engine for lodging reservations of Dutch origin. We demonstrate the results, and compare the prediction error, of several classification and regression techniques using aggregated customer reviews as a source of input.

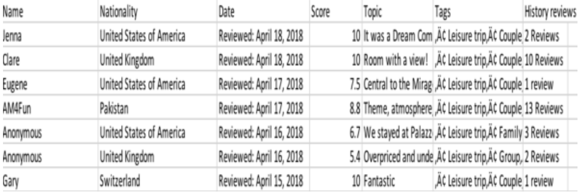
1.Introduction

Online surveys have been broadly utilized for sentiment analysis tasks. In this project, we address the rating prediction issue, utilizing sentimental analysis and topic modeling on the dataset. A rating score, in most cases, concurs with its review sentiment, which makes sentiment-words a reasonable solution for this task. To understand the dataset, we employed several APIs. We used genism, NLTK, scikit-learn, wordcloud, SpaCy.

2.Data Gathering

The data is scraped from 3 hotels' reviews published on booking.com. The hotels are Wynn Hotel, The Venetian Hotel and Caesars Palace Hotel, which are the most popular ones in Las Vegas. The website has already classified the negative reviews and positive reviews and we used selenium and beautiful soup to scrape the data. The whole information we have scraped are customer name, nationality, Overall reviews, trip objective, Date and review times and we have also scraped the classified negative reviews and positive reviews. Just these features we could apply to the natural language processing and machine learning algorithm to find the advantages and disadvantages of different hotels and to predict the rating of customers. The Table1 shows the main data we scraped.

Table1: Data scraped from booking.com



Name	Nationality	Date	Score	Topic	Tags	History reviews
Jenna	United States of America	Reviewed: April 18, 2018	10	It was a Dream Com	,Ä¢ Leisure trip,Ä¢ Couple	2 Reviews
Clare	United Kingdom	Reviewed: April 18, 2018	10	Room with a view!	,Ä¢ Leisure trip,Ä¢ Couple	10 Reviews
Eugene	United States of America	Reviewed: April 17, 2018	7.5	Central to the Mirag	,Ä¢ Leisure trip,Ä¢ Couple	1 review
AM4Fun	Pakistan	Reviewed: April 17, 2018	8.8	Theme, atmosphere	,Ä¢ Leisure trip,Ä¢ Couple	13 Reviews
Anonymous	United States of America	Reviewed: April 16, 2018	6.7	We stayed at Palazz	,Ä¢ Leisure trip,Ä¢ Family	3 Reviews
Anonymous	United Kingdom	Reviewed: April 16, 2018	5.4	Overpriced and unde	,Ä¢ Leisure trip,Ä¢ Group,	2 Reviews
Gary	Switzerland	Reviewed: April 15, 2018	10	Fantastic	,Ä¢ Leisure trip,Ä¢ Couple	1 review

3.Natural Language Processing

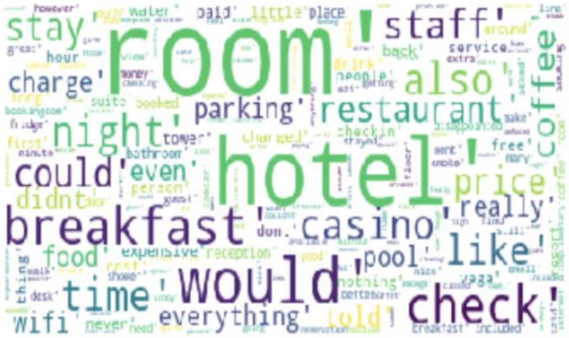
1)Data Cleaning

To clean the data, the following strategies was made:

- Using regular expression to only keep the English letters from a-z and A-Z (deleting numbers and punctuations.)
- Getting the lower case of all the words.
- Using Porter's Stemmer for stemming all the words.
- Deleting the stop words. (Stop words' dictionary from the NLTK package.)

2)Word Cloud

Word Cloud is used to find the frequency of words in the text datasets. In this section, we want to explore the disadvantages of these three hotels. Therefore, we built the word cloud based on the negative reviews.



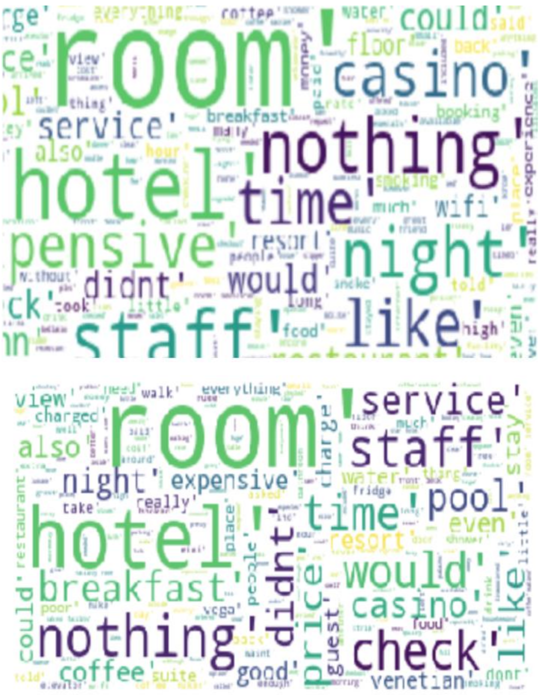
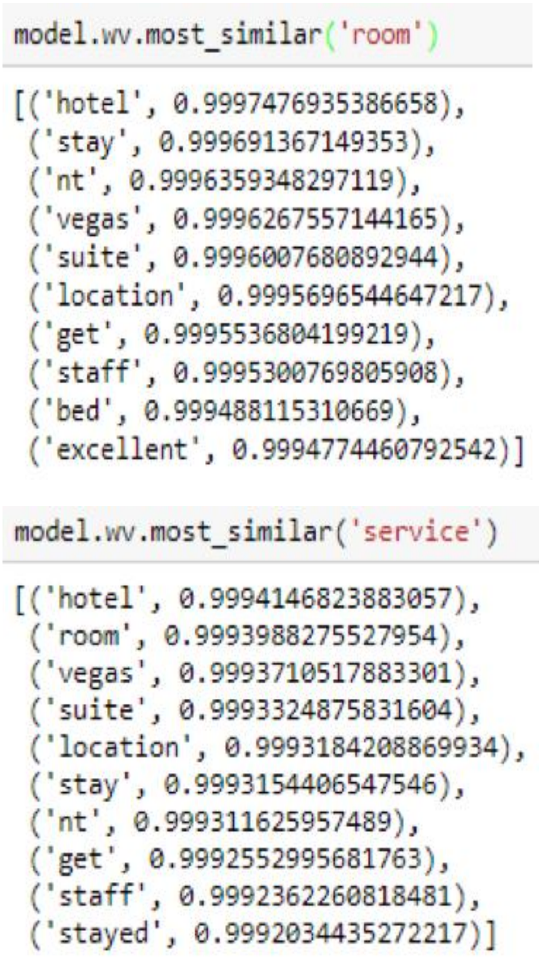


Figure1: Word cloud of the hotels(C,W,V)

In Caesars Palace Hotel, breakfast and casino are aspects customer complained. In Wynn hotel, the price and staff service are the drawbacks. In Venetian Resort Hotel, the food and service are not pretty satisfied.

2)Word2Vec

The word in the reviews need to transform to the vector and find the similarity between the words. Therefore, we could find the similarity word to what we want to explore. For example, if we want to explore the aspects of the room or service in Venetian Hotel, this model could be applied to find most relevant information.



```
model.wv.most_similar('room')
```

```
[('hotel', 0.9997476935386658),
 ('stay', 0.999691367149353),
 ('nt', 0.9996359348297119),
 ('vegas', 0.9996267557144165),
 ('suite', 0.9996007680892944),
 ('location', 0.9995696544647217),
 ('get', 0.9995536804199219),
 ('staff', 0.9995300769805908),
 ('bed', 0.999488115310669),
 ('excellent', 0.9994774460792542)]
```

```
model.wv.most_similar('service')
```

```
[('hotel', 0.9994146823883057),
 ('room', 0.9993988275527954),
 ('vegas', 0.9993710517883301),
 ('suite', 0.9993324875831604),
 ('location', 0.9993184208869934),
 ('stay', 0.9993154406547546),
 ('nt', 0.999311625957489),
 ('get', 0.9992552995681763),
 ('staff', 0.9992362260818481),
 ('stayed', 0.9992034435272217)]
```

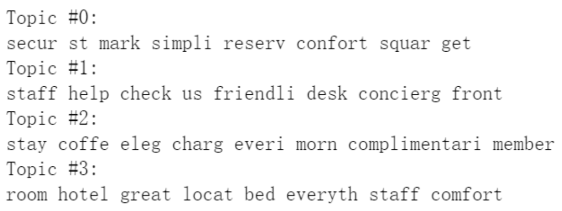
Figure2: Word2Vec Examples

3)Topic Modelling

Latent Dirichlet allocation(LDA) was used to extract topics respectively from the positive reviews and negative reviews. LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

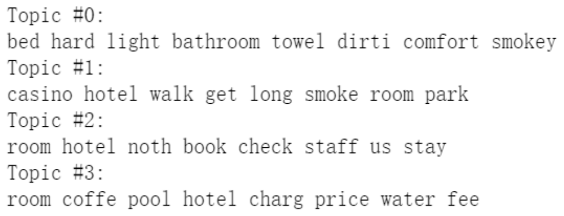
To optimize the LDA model, some parameters need to be set appropriately. After trying dozens of times, we used 4 as the numbers of topic models, 8 as the numbers of top words and no for the strategy of n-gram.

The following graphs show the results of the topic.



```
Topic #0:
secur st mark simpli reserv confort squar get
Topic #1:
staff help check us friendli desk concierg front
Topic #2:
stay coffe eleg charg everi morn complimentari member
Topic #3:
room hotel great locat bed everyth staff comfort
```

Figure3: Topics of positive reviews

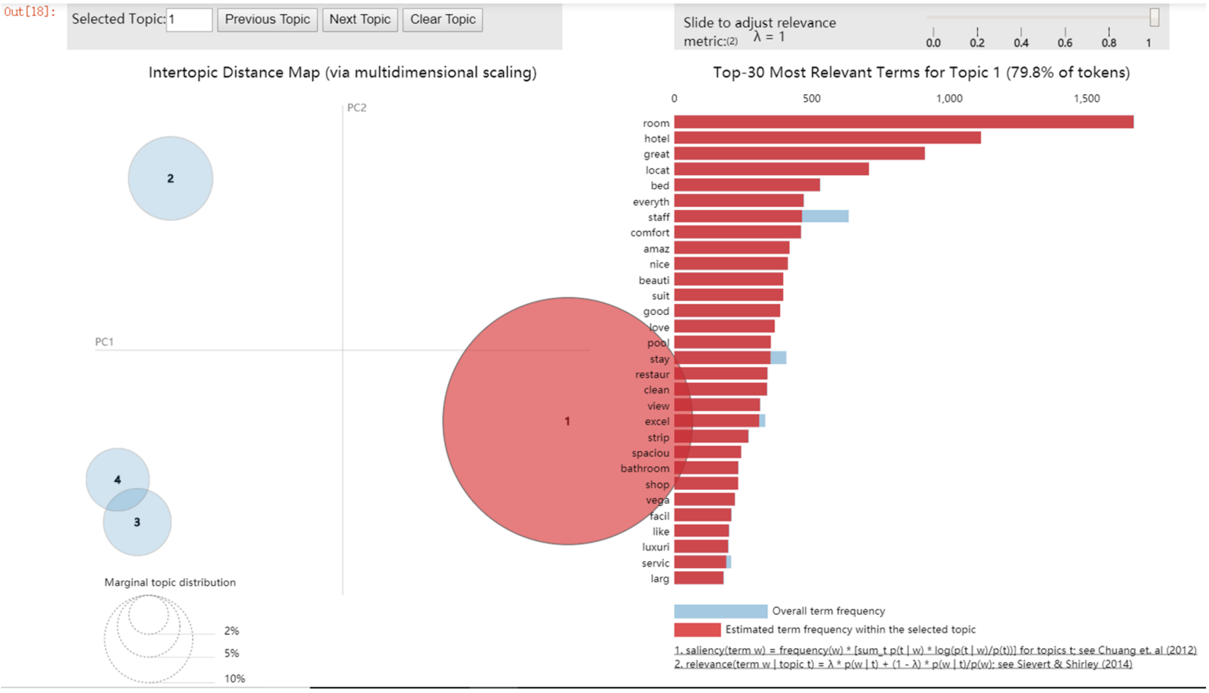


```
Topic #0:
bed hard light bathroom towel dirti comfort smokey
Topic #1:
casino hotel walk get long smoke room park
Topic #2:
room hotel noth book check staff us stay
Topic #3:
room coffe pool hotel charg price water fee
```

Figure4: Topics of negative reviews

From the topics of the positive reviews, the topics can be concluded as security and reservation process, friendly staffs, complimentary stays and coffee, and comfortable rooms and great locations.

From the topics of the negative reviews, the topics can be concluded as dirty bathrooms, long ways to casino, bad checking experience, and bad coffee and some extra fees.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure5: LDA visualization of positive reviews

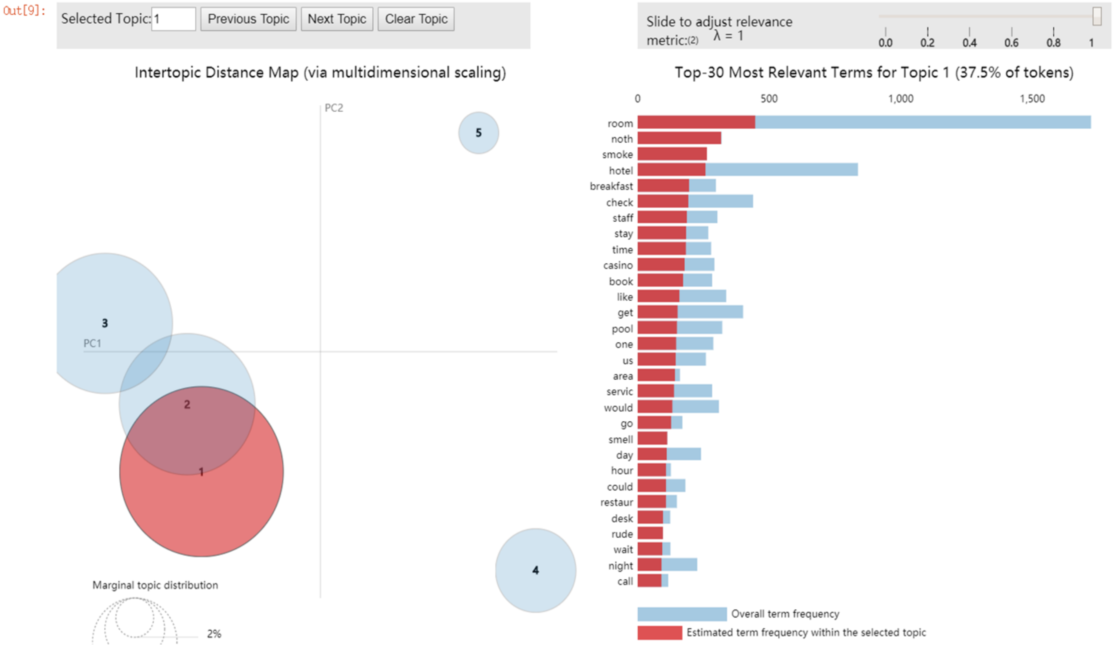


Figure6: LDA visualization of negative reviews

To further discover the topics, pyLDAvis package was used to visualize the results.

In the figure 5 and figure 6, the size of the circles represents the frequency of the topics in the documents. In this case, it means the importance of each aspect. As a result, the most positive aspect of the hotels is comfortable rooms and great locations and the negative aspects of the hotels are generally equally important.

4.Machine Learning

1) Data Preprocessing:

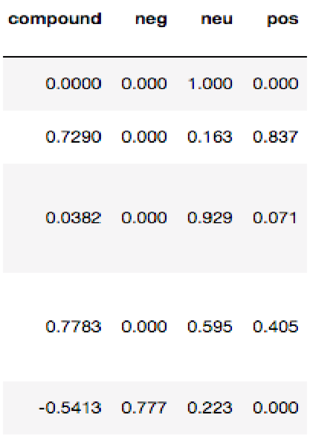
a) Increasing Features:

In our dataset, there are limited features such as Name, Topic, Date, Review times, Nationality and Tags. Then we applied the ‘get dummies’ function to get dummy variable in Nationality, Business, Family, Couple, Friends and Group.

b) Sentiment Analysis in Each Topic:

We applied the ‘sentiment intensity analyzer’ function to get the percentage of positive, negative and neutral in each topic. Also, in this way, we increase four continuous features into our original features:

Table2: Sentiment Features



compound	neg	neu	pos
0.0000	0.000	1.000	0.000
0.7290	0.000	0.163	0.837
0.0382	0.000	0.929	0.071
0.7783	0.000	0.595	0.405
-0.5413	0.777	0.223	0.000

c) The text reviews were vectorized into matrix to increase the features.

2) Exploratory Data Analysis:

a) The distribution of feature ‘score’ had been plotted and continuous feature ‘score’ was transformed to categorical data: if score greater than 8, it was labeled ‘2’; 6<score<8, it was labeled ‘1’; score<=6, it was labeled ‘0’.

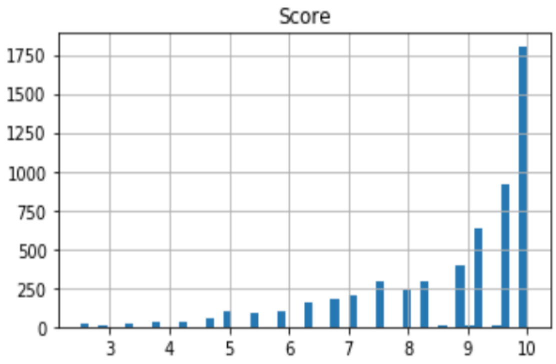


Figure7: Dependent Variable Distribution

b) We need to see the distribution of numerical data in our feature set, because our models (logistic regression, SVM) need the assumption that our data distribution is normal.

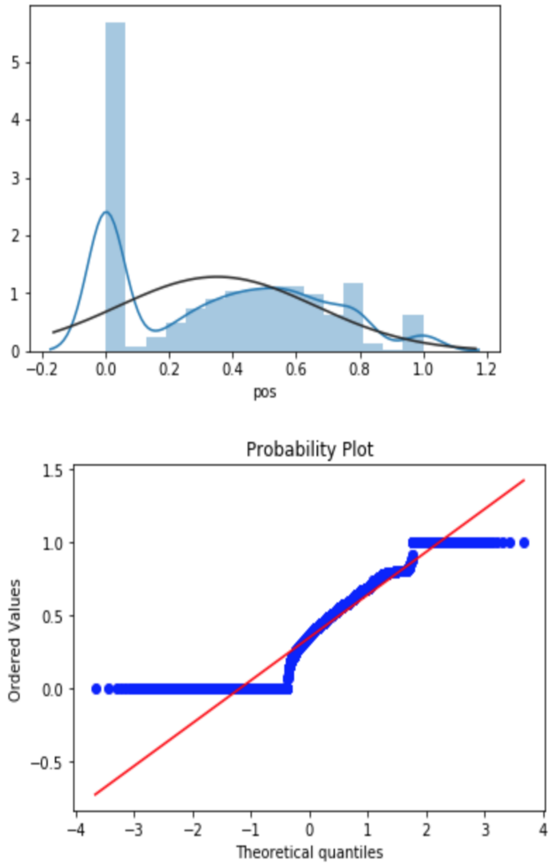


Figure8: 'Positive' Distribution

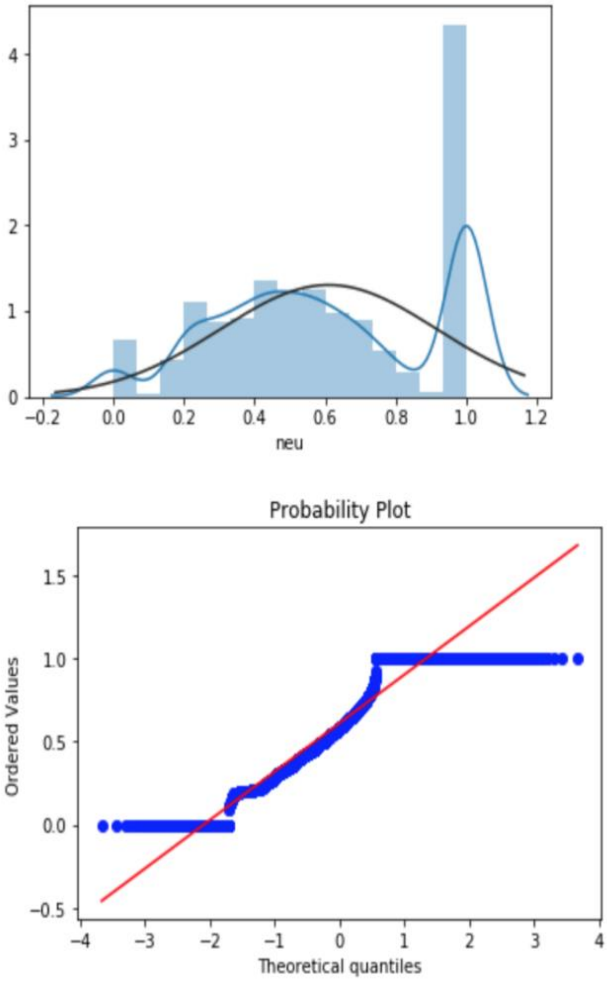


Figure9: 'Neutral' Distribution

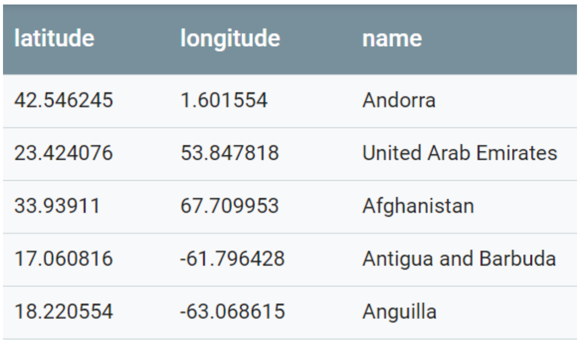
From these two plots, it is normally distributed between 0 and 1. Actually, we could use these two features into our modeling.

3)Geographical Map

To clearly visualize the relationships among the review scores, reviewers' nationality, and numbers of each country's reviewers.

For this target, we need the latitude and longitude of the countries appeared in the data. The table x shows the geographical information we used from the Internet.

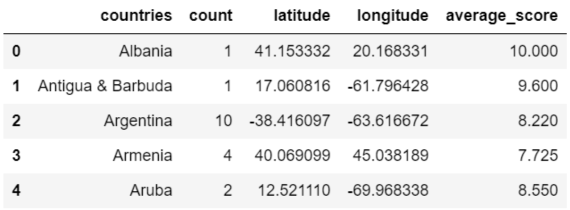
Table3: Geographical information of countries



latitude	longitude	name
42.546245	1.601554	Andorra
23.424076	53.847818	United Arab Emirates
33.93911	67.709953	Afghanistan
17.060816	-61.796428	Antigua and Barbuda
18.220554	-63.068615	Anguilla

With the geographical information and the data scraped from the booking.com, we calculate the frequency of each country as well as the average score.

Table4: Map required information



	countries	count	latitude	longitude	average_score
0	Albania	1	41.153332	20.168331	10.000
1	Antigua & Barbuda	1	17.060816	-61.796428	9.600
2	Argentina	10	-38.416097	-63.616672	8.220
3	Armenia	4	40.069099	45.038189	7.725
4	Aruba	2	12.521110	-69.968338	8.550

Finally, the data in table 4 was used to create a geographical map that shows the relationship among the review scores, reviewers' nationality, and numbers of each country's reviewers.

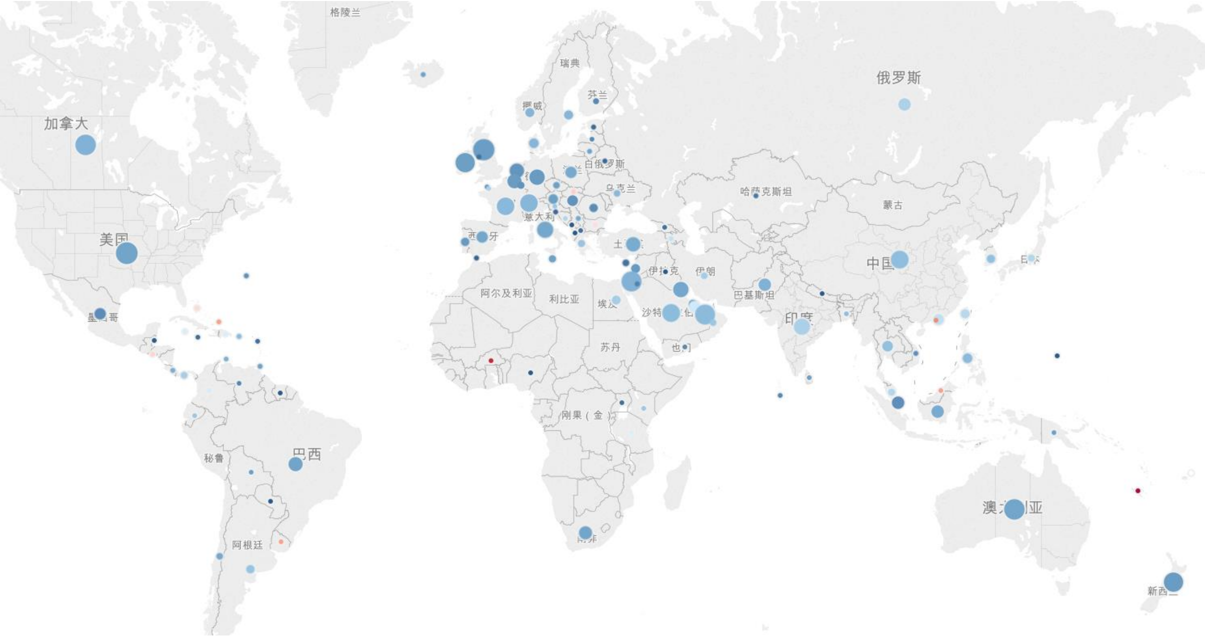


Figure10: Reviewers' map

In the map, the size of the spots represents the numbers of the reviewer and the color of the spots represents the average score of the reviews (blue means high while red means low). As a result, we can discover there exist several countries with tiny and red spot, which means the countries with few reviewers and very low score. These countries might negatively influence the performance of the model and we can try to delete these data to examine the model's accuracy.

3) Modeling:

a) Logistic Regression:

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables (Advancement Through Clarity, 2017)

When selecting the model for the logistic regression analysis, another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase the amount of variance explained in the log odds (typically expressed as R^2). However, adding more and more variables to the model can result in overfitting, which reduces the generalizability of the model beyond the data on which the model is fit.

b) Random Forest:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default) (scikit learn, unknown).

c) Support Vector Machine:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side (Petal, 2017)

d) XGboost:

GBoost is short for "Extreme Gradient Boosting", where the term "Gradient Boosting" is proposed in the paper Greedy Function Approximation: A Gradient Boosting Machine, by Friedman. XGBoost is based on this original model. This is a tutorial on gradient boosted trees, and most of the content is based on these slides by the author of xgboost.

The GBM (boosted trees) has been around for really a while, and there are a lot of materials on the topic. This tutorial tries to explain boosted trees in a self-contained and principled way using the elements of supervised learning. We think this explanation is cleaner, more formal, and motivates the variant used in xgboost (XGboost, Unknown)

4) Methodology:

Furthermore, we split the whole dataset into 80% train data and 20% test data. Using the cross-validation function with 10 kfolds, we can get the mean AUC score for the train data. GridSearch and BayesSearch could be used to tuning the hyper-parameters to obtain the optimized model and parameters.

5) Results:

Because our classification is multi-class problem, we need to transform three-class dependent variable to three dependent variables y0, y1 and y2 labeled '0' or '1'.

a) Label '2' Dependent Variable y2:

Table5: Label2 AUC Scores

Models	AUC
Logistics Regression	0.7904
Random Forest	0.7932
SVM-lin	0.7694
SVM-rbf	0.7746
XGboost	0.7633

ROC Curve:

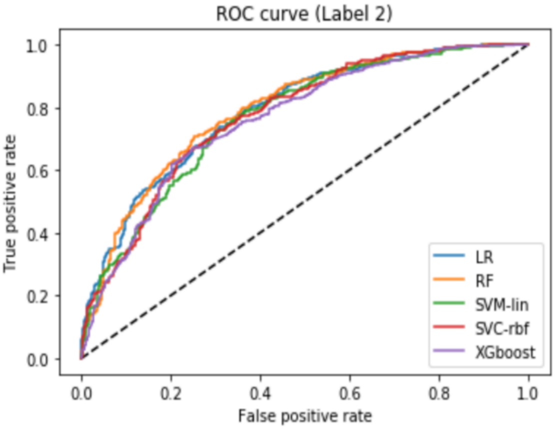


Figure11: Label 2 ROC Curve

In Label ‘2’, Random Forest algorithm has achieved the highest AUC score compared to other models.

b) Label ‘1’ Dependent Variable y1:

Table6: Label1 AUC Scores

Models	AUC
Logistics Regression	0.6916
Random Forest	0.7
XGboost	0.684

ROC Curve:

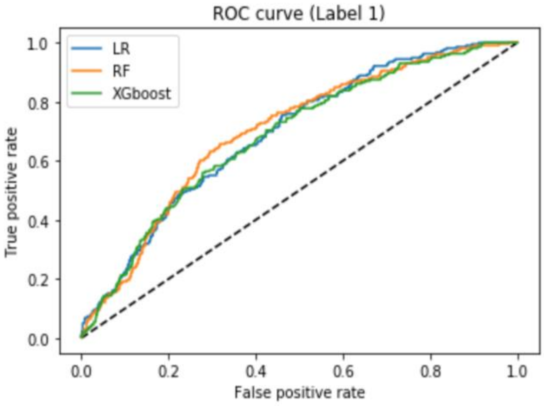


Figure12: Label 1 ROC Curve

In Label ‘1’, Random Forest algorithm has achieved the highest AUC score compared to other models.

c) Label ‘0’ Dependent Variable y0:

Table7: Label 0 AUC Scores

Models	AUC
Logistics Regression	0.7696
Random Forest	0.862
XGboost	0.85

ROC Curve:

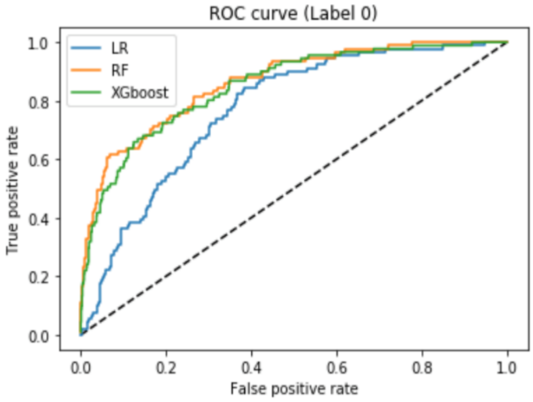


Figure13: Label 0 ROC Curve

In Label ‘0’, Random Forest algorithm still has achieved the highest AUC score compared to other models.

5.Conclusion

In our project, we have done the data scraping (data gathering), natural language processing and machine learning models. In natural language processing, the comfortable room and great location are the biggest advantage for these three hotels. But for different hotels, they have different disadvantages. For example, in Caesars Palace Hotel, breakfast and casino are aspects customer complained. In Wynn hotel, the price and staff service are the drawbacks. In Venetian Resort Hotel, the food and service are not pretty satisfied.

In the machine learning models, random forest algorithm has the outstanding performance among these three classes prediction.

However, We only used three hotels to train this model. In the future we can build a system to train the model continuously as long as new reviews are made to polish the model constantly. Besides that, XGBoost advanced model does not have outstanding performance, we would to try to fix that in the future.

6.Reference

Advancement Through Clarity (2017). 'What is Logistic Regression', Statistics Solutions [Online]. Available at: https://www.statisticssolutions.com/what-is-logistic-regression/ [Accessed 8 May 2018].

Petal, S. (2017) 'SVM (Support Vector Machine)- Theory', Machine Learning 101 [Online]. Available at:https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [Accessed 8 May 2018]

scikit learn (Unknown) 'Random Forest Classifier', sklearn.ensemble [Online]. Available at:http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [Accessed 8 May 2018]

XGboost Document (Unknown) 'Introduction to Boosted Trees', XGboost [Online]. Available at: http://xgboost.readthedocs.io/en/latest/model.html [Accessed 8 May 2018]