Groupon Review Analysis

Author: Haitao Liu, Tianyu Liu, Jhao-Han Chen, Lingyao Kong

Motivation and Objectives:

What are the major concerns of customers for a wine bar? Food? Service? Price? Can others' review affect your bar choice? In this report, we are going to unveil those essential features behind all kinds of bars using EDA, Sentiment Analysis, Topic Modeling on Groupon data. Groupon is an American worldwide e-commerce marketplace connecting customers with the cost-saving local merchants. They offer the coupons for some foods, service, travel, activities at a substantial discount in many markets. Customers can write reviews, providing a star-rating along with open-ended comment after they use their Groupon. Base on that, the reviews can have effects on consumers' choices and the merchant's reputation. For instance, a one-star increase led to a 59% increase in revenue of independent restaurants. With the rapid growth of buyers and users, we see a great potential of Groupon restaurant reviews as a valuable insights repository. In this project, we are focus on the wine bar in New York City. We want to find out the strength and weakness for each wine bar in New York City. Also, we try to identify that what aspects they need to be improved and give their suggestions.

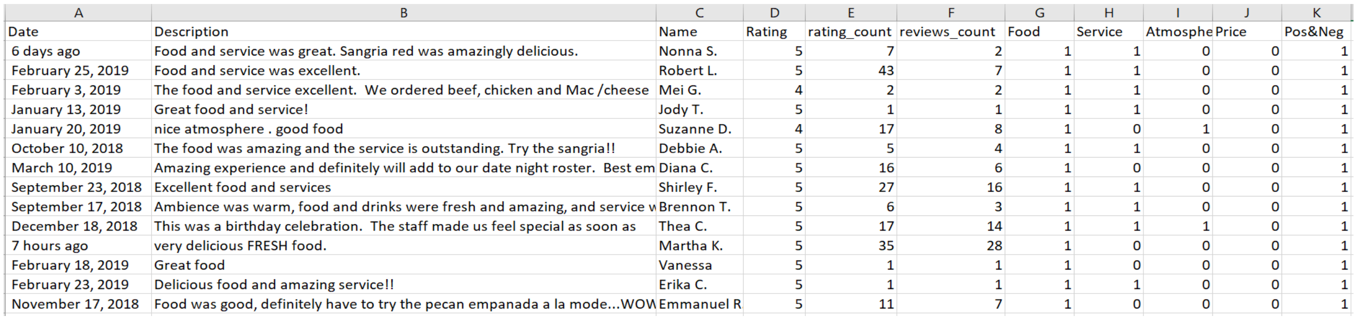
Introduction:

Sentiment analytics harnesses the power of machine learning and AI to determine whether a piece of writing is positive, negative or neutral. It helps in extracting value from the wealth of information available in the form of tweets, reviews, comments etc.Since reviews make up the greatest component for Groupon, investigations into them via machine learning techniques were expected to yield interesting discoveries. First of all, we scraped the data of 30 wine bars from the Groupon website, then applied natural language processing (NLP) to process data. Secondly, we did the Exploratory Data Analysis to find out the relationship between the rating score, rating count and reviews. Thirdly, we focused on the field of sentiment analysis which was conducted by three models, Naïve Sentiment Analysis, Support Vector Machine (SVM), Convolutional Neural Network (CNN). Then we did the topic modeling for four categories, food, service, atmosphere, price using multi-labeled classification and CNN. Last, we combined the result of the sentiment analysis and the result of the topic modeling to evaluate the wine bar and gave them some suggestions for improvement.

Data understanding:

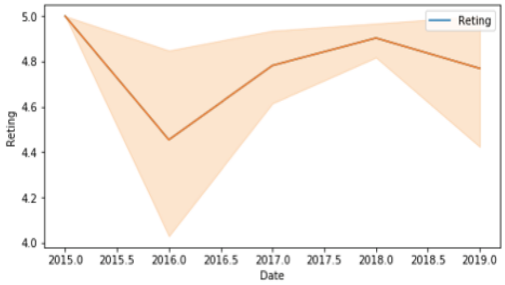
We scraped the data of 30 wine bars in NYC from the Groupon website. Stored each wine bar in CSV file, so there are 9000 reviews in total. Attributes in review data include date, description, name, rating, rating count, reviews count. We manually labeled the sentiment of the reviews for the first 6 wine bar as the ground truth. Also, we manually labeled four aspects that might be mentioned in the reviews including food, service, atmosphere and price.

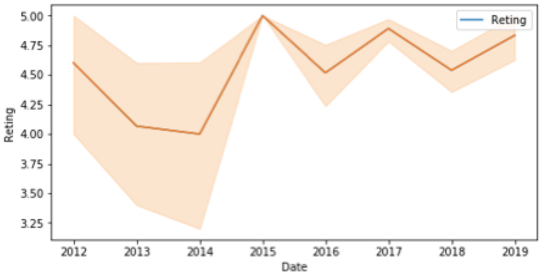
The data that prepared for the training is shown as follow:



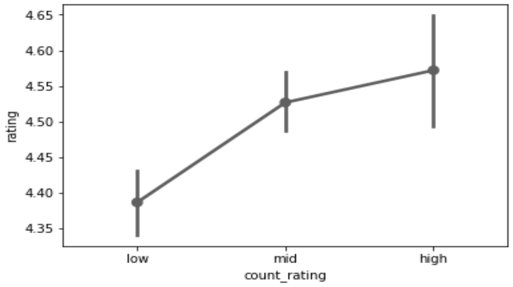
A	B	C	D	E	F	G	H	I	J	K
Date	Description	Name	Rating	rating_count	reviews_count	Food	Service	Atmosphe	Price	Pos&Neg
6 days ago	Food and service was great. Sangria red was amazingly delicious.	Nonna S.	5	7	2	1	1	0	0	1
February 25, 2019	Food and service was excellent.	Robert L.	5	43	7	1	1	0	0	1
February 3, 2019	The food and service excellent. We ordered beef, chicken and Mac /cheese	Mei G.	4	2	2	1	1	0	0	1
January 13, 2019	Great food and service!	Jody T.	5	1	1	1	1	0	0	1
January 20, 2019	nice atmosphere . good food	Suzanne D.	4	17	8	1	0	1	0	1
October 10, 2018	The food was amazing and the service is outstanding. Try the sangria!!	Debbie A.	5	5	4	1	1	0	0	1
March 10, 2019	Amazing experience and definitely will add to our date night roster. Best em	Diana C.	5	16	6	1	0	0	0	1
September 23, 2018	Excellent food and services	Shirley F.	5	27	16	1	1	0	0	1
September 17, 2018	Ambience was warm, food and drinks were fresh and amazing, and service w	Brennon T.	5	6	3	1	1	0	0	1
December 18, 2018	This was a birthday celebration. The staff made us feel special as soon as	Thea C.	5	17	14	1	1	1	0	1
7 hours ago	very delicious FRESH food.	Martha K.	5	35	28	1	0	0	0	1
February 18, 2019	Great food	Vanessa	5	1	1	1	0	0	0	1
February 23, 2019	Delicious food and amazing service!!	Erika C.	5	1	1	1	1	0	0	1
November 17, 2018	Food was good, definitely have to try the pecan empanada a la mode...WOW	Emmanuel R	5	11	7	1	0	0	0	1

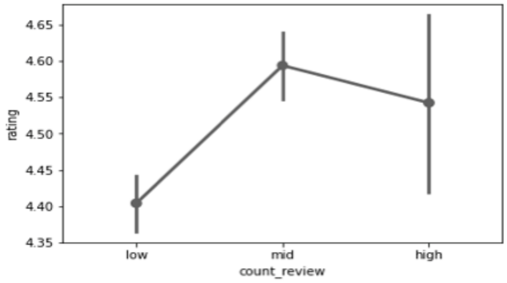
After we finished data preparation and data scraping, we also did the exploratory data analysis to analyze the dataset more intuitively. We calculated the average value of the rating by year in each store, so we can directly find the difference between the rating in each year in a certain wine bar. We plot the line chart to find the tendency for each store's average rating by year. The graphs below show the tendency of their average rating by year in 2 wine bars.





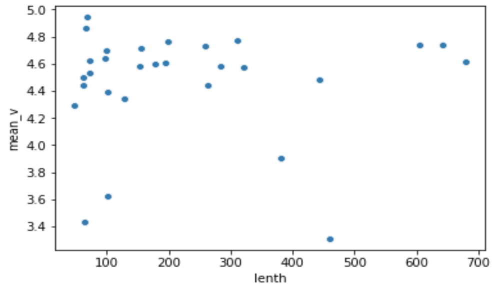
The range of orange area represents the count of rating. In 2015, there is less range of orange area because they got not less reviews, and that's the reason that the rating in this year is high. In the first graph, we can see that the rating of this store increase from 2016 to 2018,which means this store may found their weakness and improved service quality. As is shown in the second graph, the average ratings are really unstable. In general, they have a good trend over these years. Another line chart we plot is the relation between the number of reviews or rating that each people have and their rating values and the result is presented below. We classify the number of the historical rating and historical reviews for each user into 3 levels that are low(0-5),mid(5-20),high(above 20).

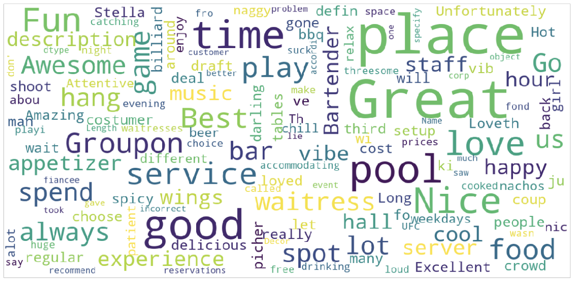




It can be seen from the line chart that people who rated more are more likely to give the higher score. And people with many historical reviews are more likely to give the higher score.

We want to discover the between the number of the rating and their rating score ,and we tried to get the more frequent word among these wine bar. So we plot a scatter graph and we did the word cloud.





It can be seen from the scatter plot that the total number of the rating is close the 200 to 300 for each bar. There is no significant relation between the number of the rating and their rating score.

Naïve Sentiment Analysis:

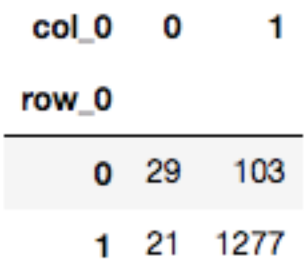
1. The model

   There are two dictionaries that contain positive words and negative words. We tried to use these 2 dictionaries to filter out the positive and negative words. Then, the number of the positive and negative words in each reviews can be counted. So the class can be assigned like majority vote:

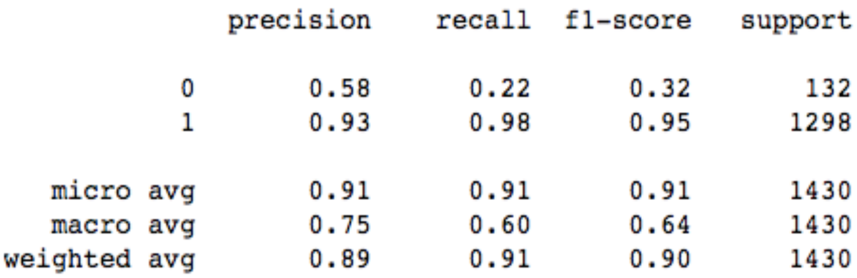
   Positive words > Negative words : Assign positive sentiment to this reviews
   Positive words < Negative words : Assign negative sentiment to this reviews
   Positive words = Negative words : Assign neutral sentiment to this reviews

2. Performance

We compared the output with the sentiment labels that we were manually labeled. We only labeled positive and negative sentiment, so the output of neutral sentiment need to be converted to either positive or negative.
If we convert the neutral sentiment to positive sentiment, the result is shown as follow:

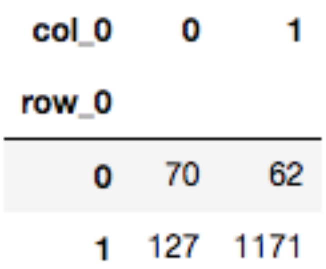


col_0	0	1
row_0		
0	29	103
1	21	1277

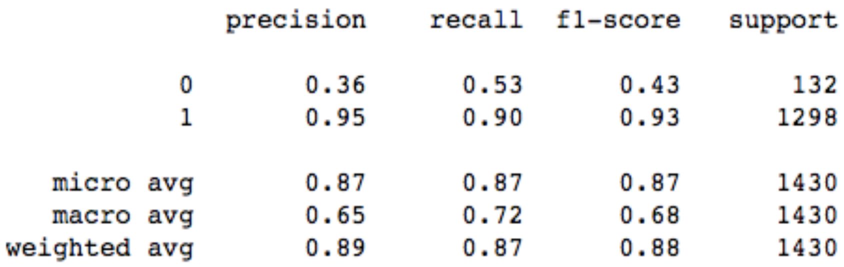


```
             precision    recall  f1-score   support

          0       0.58      0.22      0.32       132
          1       0.93      0.98      0.95      1298

  micro avg       0.91      0.91      0.91      1430
  macro avg       0.75      0.60      0.64      1430
weighted avg       0.89      0.91      0.90      1430
```

The accuracy of this model is 0.91, but the recall for negative sentiment is very low. This model nearly cannot classify the negative sentiment correctly.
If we convert the neutral sentiment to negative sentiment, the result is shown as follow:



col_0	0	1
row_0		
0	70	62
1	127	1171



```
              precision    recall  f1-score   support

           0       0.36      0.53      0.43       132
           1       0.95      0.90      0.93      1298

   micro avg       0.87      0.87      0.87      1430
   macro avg       0.65      0.72      0.68      1430
weighted avg       0.89      0.87      0.88      1430
```

The accuracy of this model is 0.87, and the recall for negative sentiment turns out better than the last one. Even though the accuracy is lower than the last one, this model can distinguish more than half of the reviews that are negative sentiment.

This project prefer to choose the second way to classify data because the negative reviews is kind of important. The main idea is to give the suggestion to the service provider, so we would love the find out negative sentiment as many as possible.

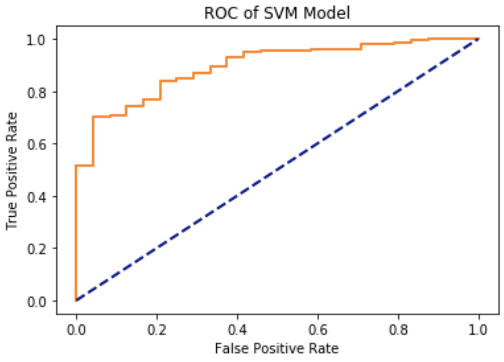
Support Vector Machine for Sentiment Analysis:

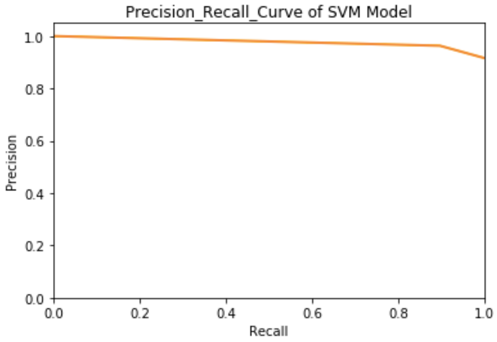
1. The Model

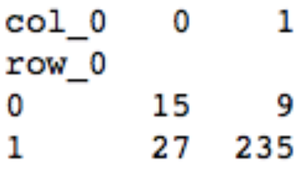
We used TF-IDF matrix as input and manually labelled sentiment as the ground truth to perform the classification. We were aim to find the sentiment for each review. We separated the data into 70% for training and 30% for testing. For the training part, 5-fold validation was applied on the training data. The original data is biased, which contain 90% of positive reviews and 10% of negative reviews. The class weight we set in this model is 1:60 because we want to distinguish as many real negative reviews as possible. The main idea in this project is to give the suggestion to the service provider. So the negative reviews are way more important than the positive reviews.

2. Performance

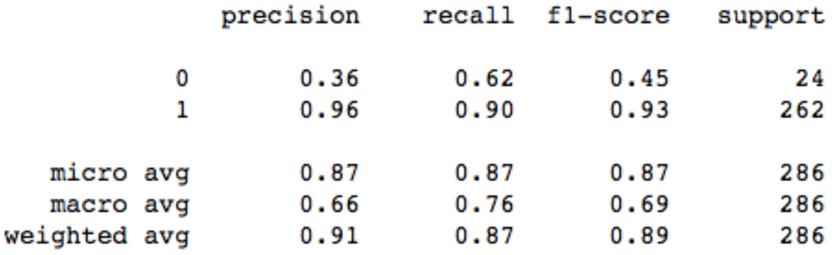
The ROC curve and precision recall curve is shown as follow:







```
col_0    0    1
row_0
0       15    9
1       27  235
```



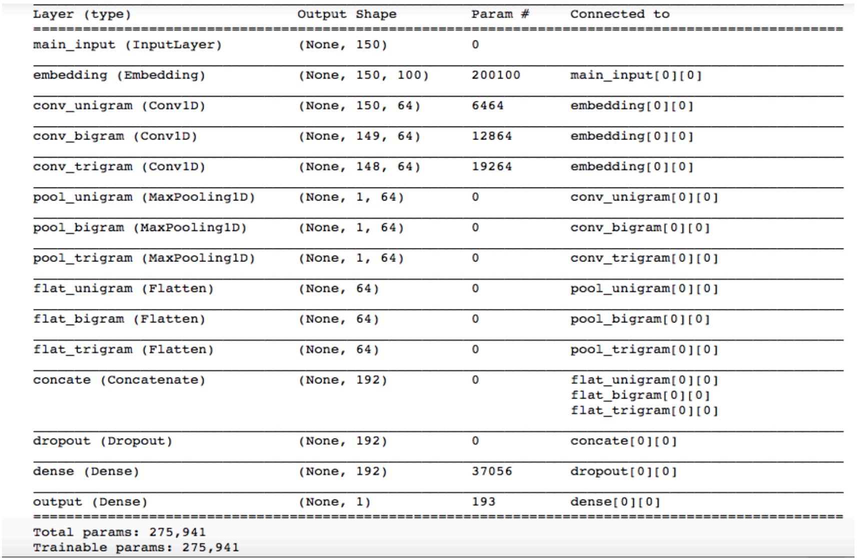
```
              precision    recall  f1-score   support

           0       0.36      0.62      0.45        24
           1       0.96      0.90      0.93       262

   micro avg       0.87      0.87      0.87       286
   macro avg       0.66      0.76      0.69       286
weighted avg       0.91      0.87      0.89       286
```

It can be seen from the ROC curve that this model have a great accuracy. The precision for negative sentiment is relatively low and the ability to predict the positive sentiment is quite good. The average AUC value of 5 fold cross validation is 0.86 and the accuracy of this model is 0.87, which are also good.

Convolutional Neural Network:

1. The Model

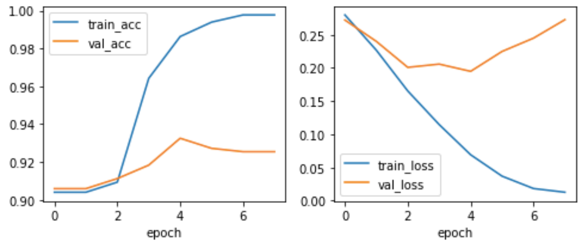
Word Embedding: I converted the sentence to word vectors, so the input is a matrix (100*150). Also, I defined the maximum words that we use is 1000. In convolutional layer, I set 1,2 and 3 as the size of the filter, which represent the unigram, bigram and trigram. There are 64 filters that we used in our model and the activation function I chose is sigmoid. In pooling layer, Max pooling method gave a best performance. We chose this method for reducing dimension and preparing for the dense layer. Then we concatenate the different outcome of filter and dropped 30% of the results. The specific summary of the model as follow:



```
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
main_input (InputLayer)         (None, 150)          0
__________________________________________________________________________________________________
embedding (Embedding)           (None, 150, 100)     200100      main_input[0][0]
__________________________________________________________________________________________________
conv_unigram (Conv1D)           (None, 150, 64)      6464        embedding[0][0]
__________________________________________________________________________________________________
conv_bigram (Conv1D)            (None, 149, 64)      12864       embedding[0][0]
__________________________________________________________________________________________________
conv_trigram (Conv1D)           (None, 148, 64)      19264       embedding[0][0]
__________________________________________________________________________________________________
pool_unigram (MaxPooling1D)     (None, 1, 64)        0           conv_unigram[0][0]
__________________________________________________________________________________________________
pool_bigram (MaxPooling1D)      (None, 1, 64)        0           conv_bigram[0][0]
__________________________________________________________________________________________________
pool_trigram (MaxPooling1D)     (None, 1, 64)        0           conv_trigram[0][0]
__________________________________________________________________________________________________
flat_unigram (Flatten)          (None, 64)           0           pool_unigram[0][0]
__________________________________________________________________________________________________
flat_bigram (Flatten)           (None, 64)           0           pool_bigram[0][0]
__________________________________________________________________________________________________
flat_trigram (Flatten)          (None, 64)           0           pool_trigram[0][0]
__________________________________________________________________________________________________
concate (Concatenate)           (None, 192)          0           flat_unigram[0][0]
                                                                 flat_bigram[0][0]
                                                                 flat_trigram[0][0]
__________________________________________________________________________________________________
dropout (Dropout)               (None, 192)          0           concate[0][0]
__________________________________________________________________________________________________
dense (Dense)                   (None, 192)          37056       dropout[0][0]
__________________________________________________________________________________________________
output (Dense)                  (None, 1)            193         dense[0][0]
==================================================================================================
Total params: 275,941
Trainable params: 275,941
```

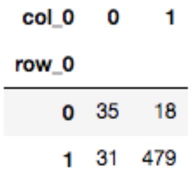
2. The Training History:

As is shown in the training history, there is a overfitting that accuracy of test data go down. So I set an early stop to prevent the overfitting. After we evaluated the model, the accuracy is around 92.3%.

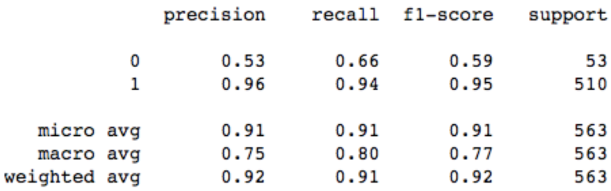


3. Performance:

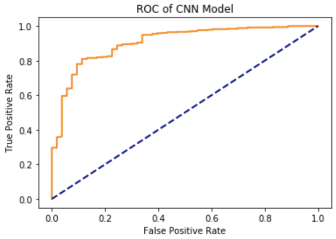
I set 0.9 as the threshold to classify the output of probability. The confusion matrix and performance are shown as follow:



col_0	0	1
row_0		
0	35	18
1	31	479



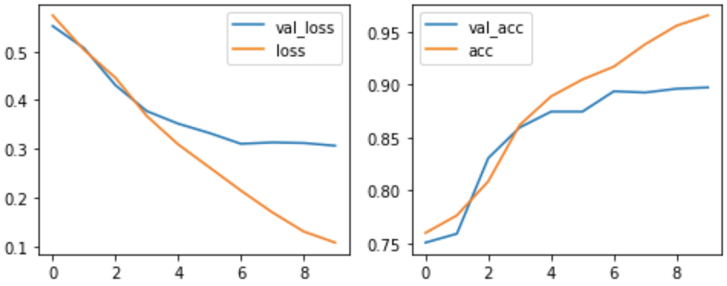
```
              precision    recall  f1-score   support

           0       0.53      0.66      0.59        53
           1       0.96      0.94      0.95       510

   micro avg       0.91      0.91      0.91       563
   macro avg       0.75      0.80      0.77       563
weighted avg       0.92      0.91      0.92       563
```



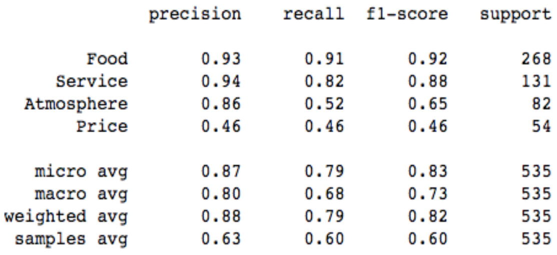
This model has good ability to predict the positive sentiment and the ability to predict the negative sentiment is also kind of good. In general, this model has a very good performance in predicting the sentiment .

4. Topic Modeling

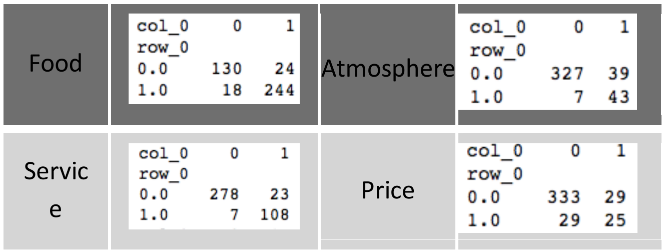
   We changed the label of sentiment to the label of aspects in the CNN model in order to predicting the aspects. We selected 4 aspects including food, atmosphere, price and service. I set the epoch value equal to 10, and here is the result of training process:



The output of probability can be classified to 0 and 1, but we assigned different threshold to different aspect(food 0.5, atmosphere 0.5, price 0.25 and service 0.15). This is because we tried to optimum the F1 score for each aspect, and the aspect like service may be biased.



```
              precision    recall  f1-score   support

        Food       0.93      0.91      0.92       268
     Service       0.94      0.82      0.88       131
  Atmosphere       0.86      0.52      0.65        82
       Price       0.46      0.46      0.46        54

   micro avg       0.87      0.79      0.83       535
   macro avg       0.80      0.68      0.73       535
weighted avg       0.88      0.79      0.82       535
 samples avg       0.63      0.60      0.60       535
```



Food

```
col_0     0     1
row_0
0.0     130    24
1.0      18   244
```

Atmosphere

```
col_0     0     1
row_0
0.0     327    39
1.0       7    43
```

Servic

e

```
col_0     0     1
row_0
0.0     278    23
1.0       7   108
```

Price

```
col_0     0     1
row_0
0.0     333    29
1.0      29    25
```

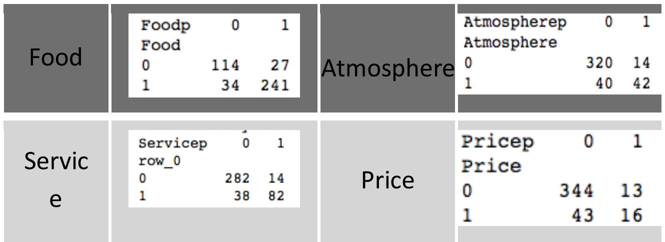
Support Vector Machine for Topic Modeling:

1. The Model

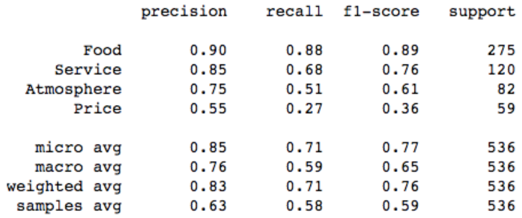
It is not enough to know the sentiment of each reviews, Knowing the topic that reviews mentioned is also important. We tried to predict 4 aspects that are food, service, atmosphere and price, so we performed the multi-labeled classification using SVM. To deal with the multi-labeled class, we used Binary Relevance that treat each label as a binary classification task. We used the same input with sentiment analysis which is tf-idf matrix and dependent variables are manually labeled 4 aspects.

2. Performance

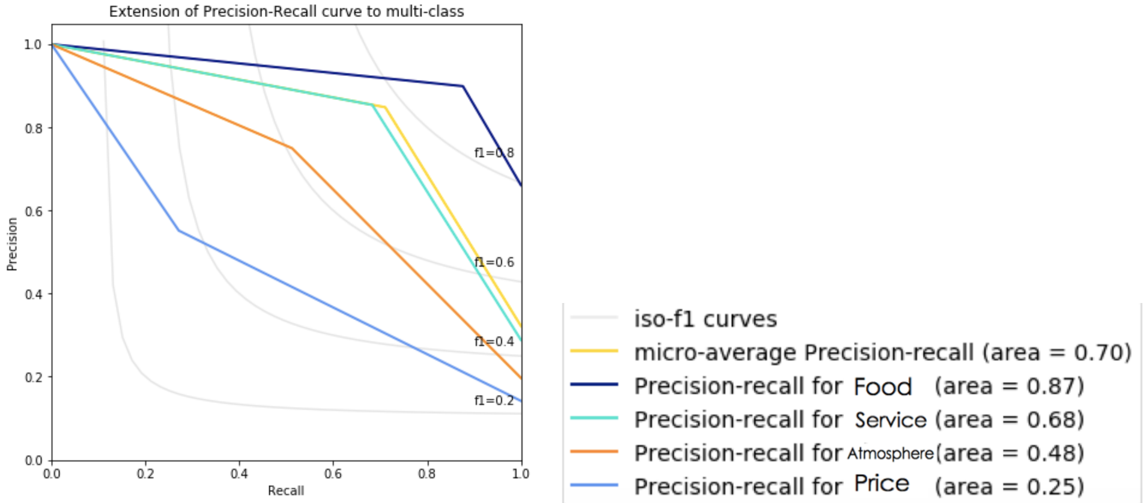
The confusion matrix for 4 aspects and their precision- recall shown as below:



Food	Foodp 0 1 Food 0 114 27 1 34 241	Atmosphere	Atmospherep 0 1 Atmosphere 0 320 14 1 40 42
Service	Servicep 0 1 row_0 0 282 14 1 38 82	Price	Pricep 0 1 Price 0 344 13 1 43 16



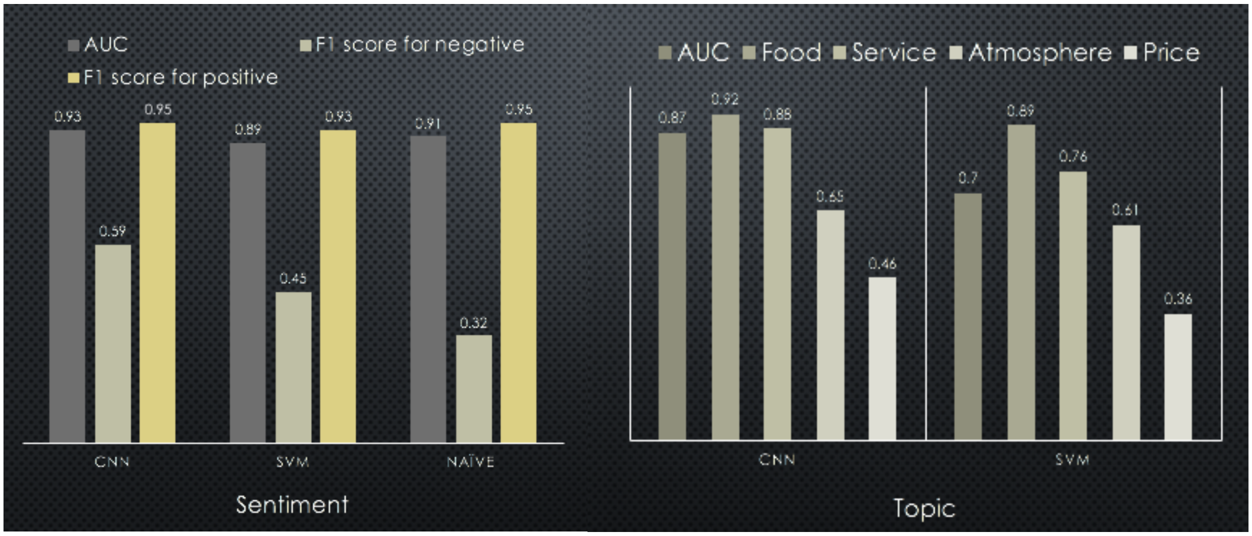
```
              precision    recall  f1-score   support

        Food       0.90      0.88      0.89       275
     Service       0.85      0.68      0.76       120
  Atmosphere       0.75      0.51      0.61        82
       Price       0.55      0.27      0.36        59

   micro avg       0.85      0.71      0.77       536
   macro avg       0.76      0.59      0.65       536
weighted avg       0.83      0.71      0.76       536
 samples avg       0.63      0.58      0.59       536
```



The precision and recall in the table are for the label 1. We want to focus on the numbers of reviews that can be correctly classified. In this model, we can find out that most of the reviews contain food and service can be correctly classified. But this model have a low ability of classifying the Atmosphere and price. Maybe the total number of the reviews that contain price is not large enough.

Model Compare

There are 3 models for sentiment analysis and 2 models for topic extraction. Their performance is shown as follow:



It can be seen from the left bar chart that CNN model has the highest accuracy(0.93) and the accuracy of SVM is relatively low(0.89). But the SVM has the highest F1-score(0.45) for negative sentiment, compared to the F1-score of CNN for negative sentiment(0.39) and the F1 score of Naïve (0.32). We want to find as many negative sentiment as we can, so we choose CNN model to predict the sentiment for the final result.

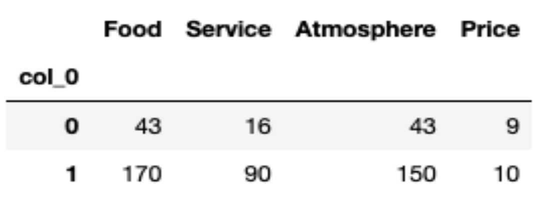
There are some other methods we were tried to use to predict the sentient, but the accuracy is not ideal enough. For example, LDA performed bad in a short text. Due to the short length of the reviews, we decided not to use this method. Also, I got a bad result to predict the sentiment using Naïve Bayes. So this method is excluded in our project.

The bar chart in the right side shows the performance of feature extraction using CNN and SVM. In general, CNN performed better than SVM in total accuracy. This chart also shows the F-1 score of each topic, only atmosphere performed better in SVM model comparing to CNN model. In all, we select CNN model to extract the topic for each review.

Analysis of Experiment results

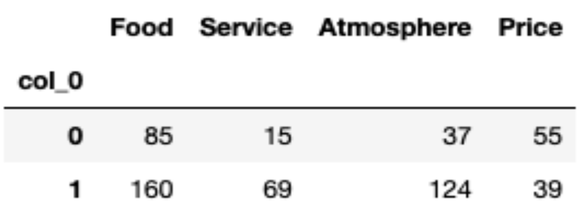
We randomly picked two wine bars to test our result. We combined the result of the SVM model and CNN model in sentiment analysis and topic modeling, and we got the matrix shown as below.

SVM



col_0	Food	Service	Atmosphere	Price
0	43	16	43	9
1	170	90	150	10

CNN



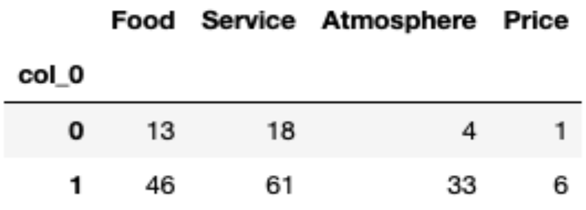
col_0	Food	Service	Atmosphere	Price
0	85	15	37	55
1	160	69	124	39

The two matrix shown above are the combined results for a wine bar called "The Wagonhouse Winery", and there are 321 reviews in total. We would like to suggest the Wagonhouse Winery to adjust their price. This is because 9 out of 19 reviews that mentioned price are negative.

Maybe atmosphere and the food in this wine bar need to be improved compared to service. The CNN model can identify more aspect of the price(94 reviews), compared to the SUM model (19 reviews). In general, the strength and the weakness that these two matrix show are similar even though we got a slightly different result.

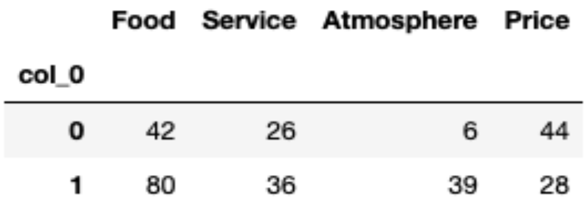
The following tables show the result of the Vinoco Wine Bar:

SVM



	Food	Service	Atmosphere	Price
col_0				
0	13	18	4	1
1	46	61	33	6

CNN



	Food	Service	Atmosphere	Price
col_0				
0	42	26	6	44
1	80	36	39	28

We can suggest that the Vinoco Wine Bar should improve their food and service in the first place and then their meal prices based on SVM model. However, the result of the CNN model shows that price is a main drawbacks. We will recommend Vinoco to take the CNN result because it has higher accuracy than SVM model.

Business Effects:

One of the most important part of running a profitable business is to attentively listen to what your customers want and how they feel so as to improve your services and increase customer satisfaction. In this report, we proposed multi-labeled classification for identifying different features for wine bar. This method was based on a high-accuracy CNN model, calculating rating score and measuring the polarity. The essential features we discovered might not only help customers to choose their favorite bars, but also provide restaurants with their advantages and shortages. On the other hand, similar procedures can be reproduced for reviews and comments in other areas like review sites, social media channels, blogs, e-commerce, etc. We can use sentiment analysis to unlock the hidden value of reviews in order to make better and more informed business decisions.

Fine tune the marketing strategy: promote its brand and services, convey the right message to the customers, listen what customers felt and thought about the brand the company was able to adjust high level messaging to meet their needs. Use topic modeling to figure out what is the competitive advantage of their products, target the right market.

Gauge the effectiveness of the marketing campaigns: the effectiveness of the marketing campaigns not only depends on the web and social media traffic but it also depends on the amount of positive discussion you are able to facilitate on social media platforms.

Improve customer service: improve service quality with max negative reviews, train employees and improve infrastructure of areas with poor service, by performing sentiment analytics the company was able to see if customers are complaining about something related to the internet or the voice services being offered by the firm and take quick corrective measures.

Future Work:

1. After making the topic table with positive and negative label, we can recommend which part the Wine bar should improve. However, when we were reading and labeling the reviews, we didn’t figure out how to label the review which has both positive and negative side of commend. For example, ‘The wine taste good! But the price is a bit expensive’ , in this review, we can’t tell if the wine taste good if we give the review a negative label. On the other side, we can’t tell if the price is not customer-friendly if we give the review a positive label. Therefore, we will try different method, e.g. Dependency parsing, to parse the reviews and detect both side of the reviews in the future.
2. In our topic table, it is not accurate enough if one topic just has few reviews. For example, the ‘price’ and ‘ atmosphere’ topic are not enough for each wine bar which just approximately contain 10% of the wine bar’s review. We will find the wine bars which have same amount of topics of reviews to train the model which is more accurate.
3. We look forward to give the sentiment prediction a score, not just a 2-levels classification. By assigning a score for each reviews, we can tell more detail about how this customer like this wine bar, but not just ‘like it’ or ‘don’t it’