Summary

November 18, 2019

1 Introduction and Background

"Yelp connects people with great local businesses." It is of great convenience for people to obtain information from the YELP app as well as share their own experience. An increasing number of business owners begin to use Yelp as an efficient feedback mechanism to achieve greater success.

In this project, we mainly focus on the businesses whose categories contain ice-cream. The optimal goal for this project is to give some specific advice to those ice-cream business owners.

2 Motivation and Statement

Our motivations in choosing models for this project basically come from following two aspects:

- **General Recommendation::** General recommendations offered by analysing properties of total businesses were based on the result of Ordered Logit Model.
  - **Ordered Logit Model:** The properties about the business are mostly categorical variables especially boolean variables. We thought that the ordered logit model is an idealt one to make such analysis.
- **Specific Recommendations:**Specific recommendations offered analysing customers review were based on the predicted score according to LSTM and Chi-square test conducted on word count dataset.
  - **Score prediction based on recurrent neural network:** We tried to predict some scores from three perspective: **flavor, price, service and environment**. Consider the scores is calculated from user's rating, we try to predict the scores in different aspects based on user's review. Under this situation, the Recurrent Neural Network especially the LSTM(Long Short-Term Memory) has a well-known performance in text analysis, which drives us to employ this methods as part of our model.
  - **Chi-square Test on positive/negative word count:** We conduct the hypothesis test on each combination of business and key word. The hypothesis test with p-value smaller than 0.05 would show significant strength or weakness in aspect related to that word.

3 Data Preprocessing

There are 2764 businesses whose category contains `icecream`. We combined the business information with reviews and obtained a total of 123397 reviews. Our data preprocessing mainly focused on two files: business and review.

We conducted text cleaning in following steps for analysis and visualization:

- **Stopwords and Special Punctuations Deletion**: “down”, “they”, “their”, “what”, “a”, “.”, “”’, “[", "]”, “(”, “)” et al.
- **Word Lemmatization**: transform nouns, verbs and adjectives form to its basic form, e.g., *criteria* to *criterion*, *went* to *go*, *better* to *good*.

We conducted data cleaning of the business JSON file in the following several steps:

- **Variable Transformation**: calculate the average opening hours weekly and transform numerical opening hours to 3 variables: Morning, Afternoon, Evening. Each variable is a categorical variable with 3-level values: True, False, and None.
- **Attributes Selection**: selecte the attributes with null values less than 1000: `RestaurantsPriceRange2`, `WiFi`, `BikeParking`, `BusinessParking`, `RestaurantsTakeOut`, `BusinessAcceptsCreditCards`,.
- **Sub-Attribute Transformation**: transfrom attributes which consists of several sub-attributes into a 3-categorical variable: `TRUE`, `FALSE`, `NONE`.

4 General Analysis

We conducted the general analysis based on three aspects: flavor, location and rating, by supposing that these three aspects are inner-related.

- The **flavor** information extracted from whether it is mentioned in certain review contains: Chocolate, strawberry, vanilla, mango, caramel, banana, coconut and raspberry, information is extracted from whether it is mentioned in certain review.
- The **location** extracted from business data contains the states of ice cream business.
- The **rating** extracted from review data contains the rating of certain review.

We made contingency tables between any of two aspects calculting the count of reviews. We conducted Pearson’s Chi-squared test to check independency. Friedan rank sum test was further conducted in case of the violation of assumption in Chi-squared test.

- **Relationship between flavors and ratings:** the Chi-squared test and Friedan rank sum test all showed p-values below 10-e6, concluding the flavors is related to rating on whole data.
- **Relationship between flavors and locations:** the Chi-squared test and Friedan rank sum test all showed p-values below 10-e13, concluding the flavors is related to the belonging states of ice cream shop.
- **Relationship between locations and ratings:** the Chi-squared test and Friedan rank sum test all showed p-values below 10-e8, concluding the states of ice cream shop is related to rating on whole data.

5 Main model

We constructed different models for two purposes mentioned before. We applied **ordered logit model** on business data to gave advice on all business regardless of review. We applied **Recurrent neural network** and **Chi-square test** to raw review text to gave advice on certain business.

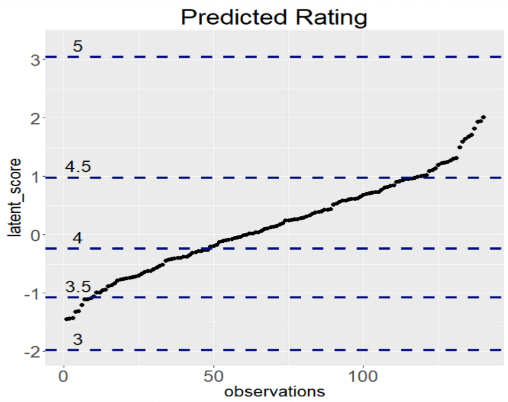
5.1 General Recommendations

5.1.1 Ordered Logit Model

We applied the Ordered Logit Model on business data. We treated stars from 0.5 to 5 as responses, while other attributes as predictors including opening hours(morning/ afternoon/ evening), whether the restaurant accepts credit cards, price range, whether the restaurant accepts take out, WiFi, whether the restaurant offers parking space for bikes and cars.

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta X, j = 1, \ldots, J - 1$$

On the right side of the equal sign, we see a simple linear model with one slope, β, and an intercept α_j that changes depending on j. Here the j is the level of an ordered category with J levels. In our case, j = 1 stands for 0.5 stars. So we see we have a different intercept depending on the level of interest. In our example, $P(Y \leq 8)$ means the probability of below 4 being 4 or above. Thus we're using the levels as boundaries.



This image shows the general businesses and their predicted ratings. We came up with the recommendations to ice cream shops. - Ensure the opening hours contains afternoon. - Provide free WiFi. - Provide parking lots. All these procedures prompt ratings to business.

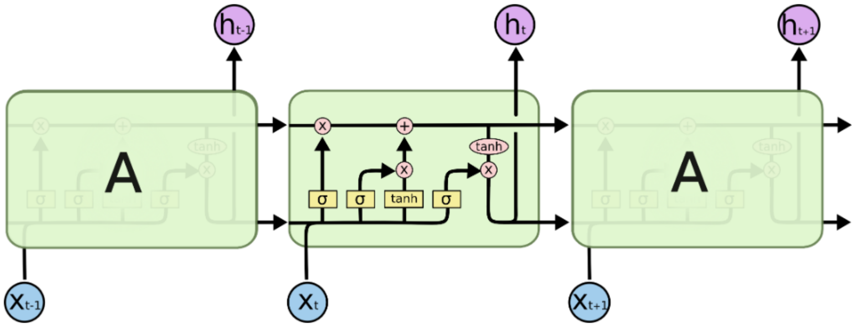
Variables	MorningTrue	AfternoonTrue	EveningTrue	AcceptCreditCardFalse	PriceRange2	TakeOutTrue	BikeParkingTrue	WifiFree	WifiNo	Parking
Coefficient	- 0.7105	1.5347	- 0.6939	1.0513	- 0.3242	0.9857	- 0.1150	0.6148	- 0.0722	- 0.2246
Standard Error	0.1117	0.2715	0.2469	0.2939	0.1170	0.2434	0.1377	0.6161	0.6132	0.1116

This image shows the general businesses and their predicted ratings. We came up with the recommendations to ice cream shops. - Ensure the opening hours contains afternoon. - Provide free WiFi. - Provide parking lots. All these procedures prompt ratings to business.

5.2 Specific Recommendations

5.2.1 Recurrent neural network

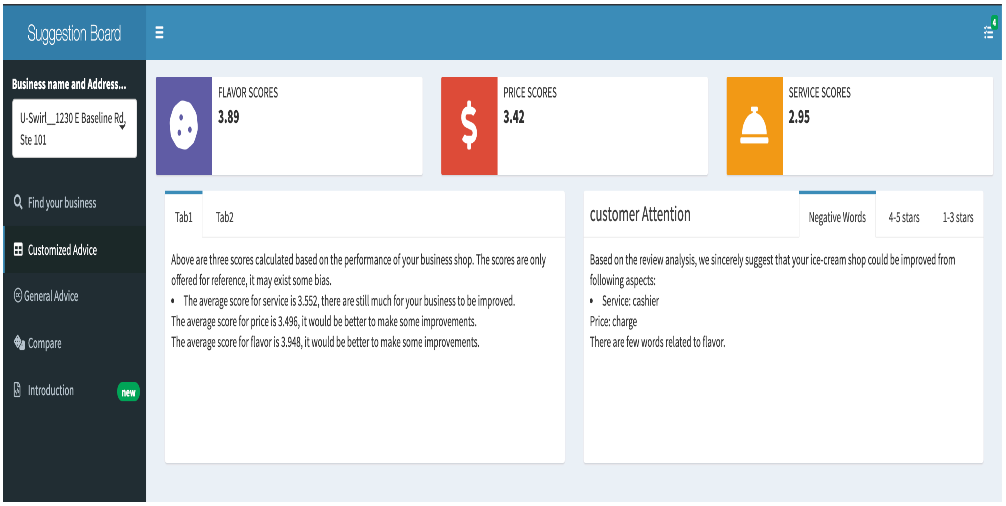
We first conducted word embedding using Global Vectors for Word Representation(GloVe). Then we constructed our model. We used Long Short-Term Memory(LSTM) combined with Bidirectional RNN. We added two hidden layers with 128 nodes in each layer. The following diagram shows the basic cell of LSTM(image source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/).



The repeating module in an LSTM contains four interacting layers.

We used the following procedure to give business owners several suggestions based on LSTM model, given reviews for a certain business: - Break all the reviews into subsentences. - Divide all subsentences into 3 categories: **food**, **price**, **service and environment**. For each category, the keywords are the same as those mentioned in word count interpretation. - Predict outcome on certain category by feeding the fitted RNN model with subsentences related to this category.

Here is one example of advice we make to the ice cream shop **So Cool Frozen Yogurt**, for more results, please refer to our shiny app.



Business name and Address...

U-Swirl__1230 E Baseline Rd, Ste 101

FLAVOR SCORES 3.89

PRICE SCORES 3.42

SERVICE SCORES 2.95

Above are three scores calculated based on the performance of your business shop. The scores are only offered for reference, it may exist some bias.

- The average score for service is 3.552, there are still much for your business to be improved.

The average score for price is 3.496, it would be better to make some improvements.

The average score for flavor is 3.948, it would be better to make some improvements.

Based on the review analysis, we sincerely suggest that your ice-cream shop could be improved from following aspects:

- Service: cashier

Price: charge

There are few words related to flavor.

5.2.2 Interpretation from word counts

We conducted a deep analysis to explore the specific aspects in which we may offer recommendations to the business owners. Intuitively, if a certain word is placed more closely with a positive word, we may think the review reflects a positive attitude towards this word. Therefore, we do following several steps to specify our result.

- Select the most frequently appeared words (especially nouns) in the word-count dataset. The selected words are shown below:
  - **Food**: taste, flavour, topping, Chocolate, strawberry, vanilla, mango, caramel, banana, coconut, raspberry.
  - **Price**: price, combo, charge.
  - **Service and Environment:** service, cashier, customer, owner, seat, table, employee, location, manager, drive, park, room, experience, atmosphere, street, staff, space, machine, decor.
- Find the appearance number of nearest positive words or negative words based on the data set [1-2]. By specifying the nearest positive/negative words towards our target words, for each target words we counted the total number and recorded it. We also performed the same counting methods within each business's review.
- Conduct the Chi-square test on 2×2 contingency table as it shown in the following table, which consists of positive and negative words counts for total reviews and a certain business.

Key Word	# of Positive	# of Negative
chocolate in Business I	34	12
chocolate in total Reviews	19629	8869

- Based on the p-values ($p\text{-value}<0.05$) and the positive/negative proportion, point out several specific areas to business owners that need to be improved. If the p-values is below the significance level, we think the attitude towards this word in certain business is different from that in general case. Then if the proportion of negative count is greater, we may think the business owner need to improve the shop in this aspect.

6 Strengthes and Weakness

6.1 Strengths

- We conducted evaluations based on three aspects of each business, which offers more comprehensive feedback to business owners.
- We evaluated the contribution of facilities and opening hours towards rating in business shops in order to offer recommendations from a different perspective.
- We roughly tested the attitude in reviews based on the word count related to adjective words. The result can demonstrate the weakness in a relatively specific aspect to business owners.

6.2 Weakness

- The score predicting methods in different aspects of the business shop may lack certain exactitude due to some overfitting problems.
- The result of ordered logit model reflects certain information which consistents to our common sense. However, the coefficients of certain features contradicted to our essential sense.
- The chi-square test on word counts dataset may be weak since the attitude is more complex than just a counting number related to adjective words. Except for the relationship between attitude and counting numbers, there may exist some business shop who have relatively fewer reviews, therefore it is hard to conduct the hypothesis test on 2×2 matrices with zeros exist.

7 Conclusion

We achieved our ultimate goal of making recommendations to the business owner in two aspects. - We proposed several hypotheses on business and validated them. - We made general recommendations to all business owners like: provide chocolate flavor, ensure opening hours contain afternoon, provide free WiFi, and provide parking lots. - We also made recommendations to certain business owners regard specific aspect **food**, **price**, **service and environment** based on predicted ratings.

We came up with two models, **Ordered Logit Model** and **Recurrent Neural Network**. - The **Ordered Logit Model** had nice interpretation on all business owners. - The **Recurrent Neural Network** had reasonable accuracy while lacking interpretations. We can make recommendations based on new predictions.

8 Reference

[1]Minqing Hu and Bing Liu. “Mining and Summarizing Customer Reviews.” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.
[2]Bing Liu, Minqing Hu and Junsheng Cheng. “Opinion Observer: Analyzing and Comparing Opinions on the Web.” Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

9 Contribution

Zhuoyan Xu: data preprocessing on reviews, construct LSTM model and Ordered Logit Model.
Rita Wu: data preprocessing on attributes, Chi-square test based on word count.
Chenghui Li: Shiny App.