

The Sentimen Evaluation of Customer Review in Las Vegas City Restaurants

Pozy Pak Ya

October 22, 2015

Abstract

*Online review becoming so important for business nowadays. Every review can affect the products buying power or services provided, including the ratings, which can quantify the satisfaction of the reviewer's experience. The source of the information mainly coming from the reviewers comments and the tip which contain resourceful textual information which been entered into the website or portal. In this paper, we evaluate the textual information from the YELP review dataset and try discover the hidden pattern or message inside it. We applied a machine learning techniques such as **Naive Bayes** to learn from the word vector which capture the sentiment from each reviewer in relation with other factor from the dataset. We mainly focus our discovery towards the reviews of the restaurant in major area of Las Vegas.*

Keywords

Urban design, social media, geo-location, lexicographic analysis, sentiment analysis

Introduction

The evaluation is about the sentiment analysis over the review and stars rating restaurant in the Last Vegas city.. The YELP dataset is very resourceful which provides the valuation criteria over 61,184 unique records for **business** , 1,569,264 records for **review** and 495,107 records for the **tips**. Two tables have been discarded for now ,which is **user** details and the **check-in** information. The GPS longitude and latitude available inside the **business** dataset which provides very useful information about its geolocation. The star value gives the feedback from the customer which might be **positive** , **negative** or **neutral**. The findings offer exemplary **big data** analysis methods as the evaluation of socially mediated urban space associated with the pattern classification of textual information inside the **reviews** and **tips** in relation with **business** dataset.

Las Vegas City is the top 5 locations with the most review counted as follows :-

Table 1: Top 5 City Reviews and Categories

| business_categories | city | review_count |
|---|-----------|--------------|
| [Breakfast & Brunch, Steakhouses, French, Restaurants] | Las Vegas | 4578 |
| [Sandwiches, Restaurants] | Las Vegas | 3984 |
| [Buffets, Restaurants] | Las Vegas | 3828 |
| [Buffets, Restaurants] | Las Vegas | 3046 |
| [American (Traditional), Restaurants] | Las Vegas | 3007 |
| [Buffets, Restaurants] | Las Vegas | 2949 |

In more details , Mon Ami Gabi is the top of 5 Las Vegas restaurant by the most reviewed counted as follows :-

Table 2: Las Vegas City Restaurant

| name | stars | review_count |
|------------------|-------|--------------|
| Mon Ami Gabi | 4 | 4578 |
| Earl of Sandwich | 4.5 | 3984 |
| Wicked Spoon | 3.5 | 3828 |
| Bacchanal Buffet | 4 | 3046 |
| Serendipity 3 | 3 | 3007 |
| The Buffet | 3.5 | 2949 |

The summary of the `joined` dataset as follows :-

Table 3: Summary of Las Vegas City Restaurant No. Of Review

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|------|
| 3 | 9 | 26 | 96 | 87 | 4578 |

To reduce the size of the sample , average size of numbers of message is the minimal size which is around 390. And the numbers of group identified around 1000

From the summary show that the Median is **26** and we choose **26** as the minimal sample for this evaluation. The median better than mean because of it is a symmetrical statistic and more resistant to errors.

Methods and Data

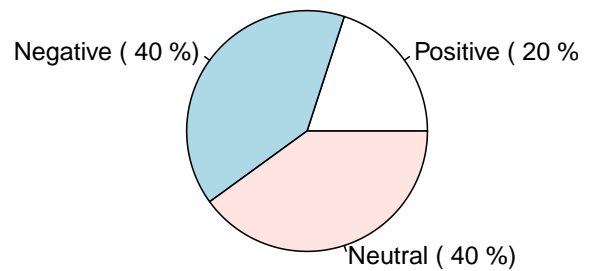
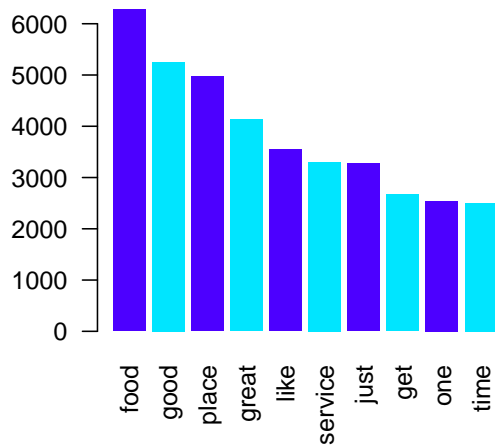
The dataset is obtained from the YELP website (http://www.yelp.com/dataset_challenge) and extracted. The format for the dataset is in `JSON` . `JSON` need special techniques to parse and read from it. Apache Hive is the best component which is capable read this format . Since the dataset required to have a good machine in term of CPU and memory , we push this dataset to work inside Hadoop which Map-Reduce can be used as the framework for the filtering and cleaning over large size of the dataset. Hive is compatible to use scripting parameter similar to `SQL` and this is very suitable for speed up the entire development work. Hive also support for the complex data type and `STRUCT` is used to handle the `JSON` complex type for the table creation inside Hive .

For the basic analysis , this evaluation requires a fair amount time to know about the dataset abd performaing exploratory analysis. But, now we only focus on the textual information which mostly inside the `review` and `tips` dataset in conjunction with the `business` and `user` information. This will tackle some of the questions such as :-

- What is the emotion type that might contain inside the review and tips messages ?
- What is the most frequent words or terms inside it ?

Below is the answers for the questions above . Top common words inside the review is regarding the **good food** , **good places** and also a **good services**. Reviewers whom visits seems very happy about the quality of food , services and places restaurant in Las Vegas. Most of the comments seems positively accepts it.

Emotion Polarity of Review Messages



- To discover what types of food been reviewed most by the reviewer , **word-cloud** plot is used to splot the word frequencies. Since to have the food-list dataset is hard to compile due to there is a lot of food around the world , by plotting it into word cloud we easily can identify manually pick the food that we recognize as follows :-



Other interesting findings in this evaluation is to classify the reviewers ratings and the tips provided. The idea is to calculate the sentiment score for each messages so we can know how positive and negative the messages. Below is the formula of the how to calculate the score :-

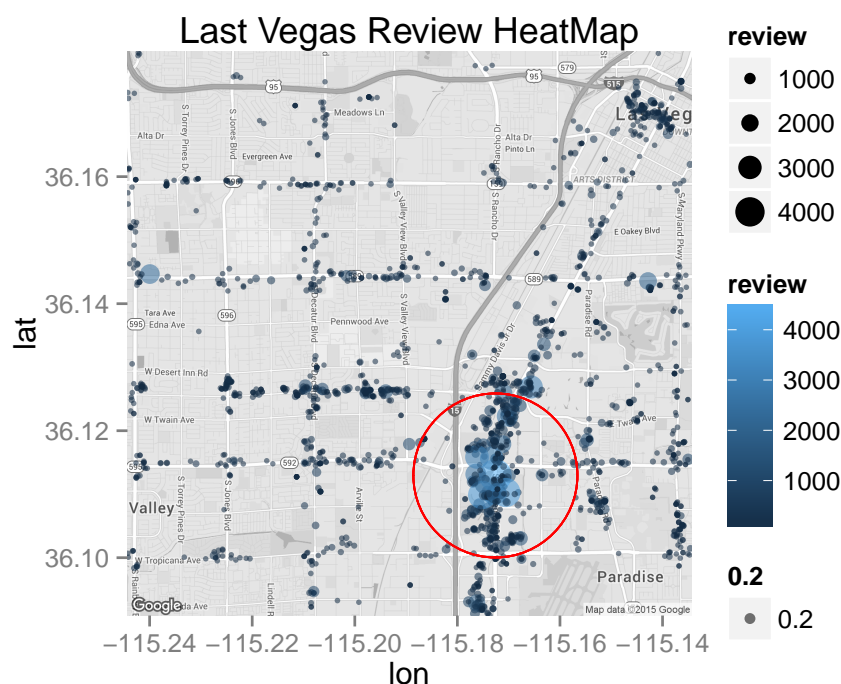
- If the `score > 0` , the messages has overall positive opinion

- If the `score < 0` , the messages has overall **negative** opinion
- If the `score = 0` , the messages has can be consider as **neutral** opinion

The lexicon is in English and the reference for the **positive** and **negative** words is reference from (https://github.com/SamPortnow/Depression_Prevention_Program/tree/master/bato/assets).

Results

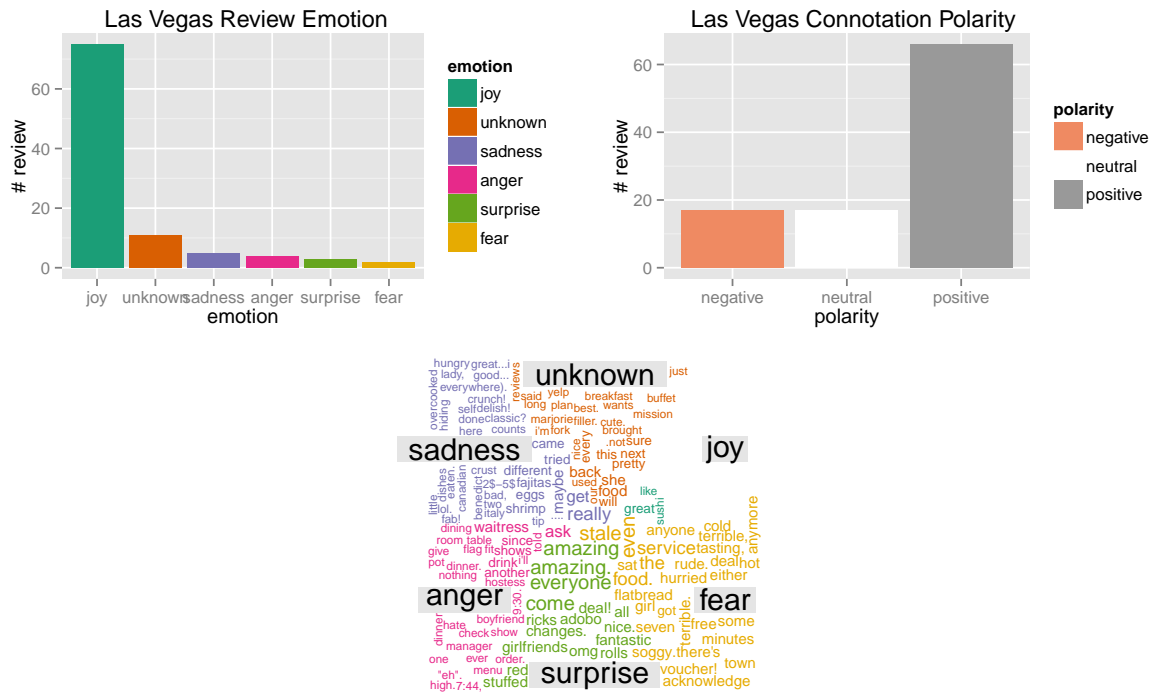
The results from the analysis we can summarize by plotting the **heatmap** of the message size inside the map of Las Vegas restaurant. We concluded that the message is more focus in the area of **Fountains of Bellagio** along the **S Las Vegas Blvd** road. This road is the main highway in Las Vegas and there is a lot of casinos along it. The illustration below shows the findings :-



The red area is the central area of the review messages. Major comments all over the year since 2004 is focused in that area. The location seems very strategic and it is only 2.5 miles from the **McCarran International Airport**. Most people will stay in this area and enjoyed the food before they have their flight or after they arrive here.

From the random sample of the Las Vegas dataset which taken 10% of the total sample for analysis, we found that more than 70% are happy (**joy**) emotion with the services, smaller portion are unhappy (**sadness**, **anger**, **fear**) emotion. The (**unknown**) emotion still needs deeper analysis to increase accuracy of the classification. This smaller sample used due to limitation of machine memory to process

A **Naive Bayes** algorithm classifies each review message into subjectivity and polarity from the input. A subjectivity indicates the words are strongly or weakly subjective and the polarity indicates the weather the words are expressing a positive or negative sentiment. The illustration below shows the summary:-



Discussion

The issue of using such real data in this situation raises several significant questions such as :-

- 1) How to determine the extent of limited demographic information, data frequency, the privacy concerns of Yelp users and harsh message that might be censored by YELP administration.
- 2) Accessing and processing method due to increasingly for larger datasets. Some algorithm like Random Forest required high iteration and processing.
- 3) Assuring a sample's diversity by implementing feature selection and optimal ratio selection for training and testing dataset for prediction in the future.
- 4) Authentication or validation of the reviewer's identity, to reduce or eliminate bias in the reviewer comments so all the information is genuinely entered and the analysis result can be trusted.

We believe that the three analysis techniques can be improved by suggestion below :-

- 1) How Big Data can help the analysis as a robust tool to support the high volume dataset for better analysis results.
- 2) How the selection of strategic location can be correlated with the review pattern in order to reveal the dynamic sentiment in social trends.
- 3) How the star ratings provided can be mapped to the analysis, which allows to make the analysis model is entirely in qualitative measure. By adding the polarity between the positive and negative also can improve the model.

Textual reviews dataset in YELP dataset for different products and services is still abundant. The task needed in getting sufficient and reliable information can become a daunting task. In the future, we will incorporate to include the analysis using both classification and regression model for better predictive analytics. We might find a better and stronger interdependence in the way of the user ratings in different aspects. Finally, the current evaluation and the analysis are reproducible due to the dataset is available to download and all the source code which involved is publicly shared via the GitHub website (https://github.com/pozypakya/capstone_yelt.git)