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 areas of Las Vegas.

Keywords

Social media, geo-location, sentiment analysis

Introduction

The evaluation is about the sentiment analysis over the YELP reviews and tips about restaurants 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 in the `business` dataset provides very useful information about its geo-location. The reviews and tips value contains `positive`, negative or neutral feedback from the customer.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 the `business` dataset.

Which city is the top 5 locations with the most review counted ?

Table 1: Top 5 City Reviews and Categories - Las Vegas

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

Which restaurant is the most reviewed counted ?

Table 2: Top 5 Las Vegas City Restaurant - Mon Ami Gabi

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 the message is the minimal size which is around 390. And the numbers of the 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.

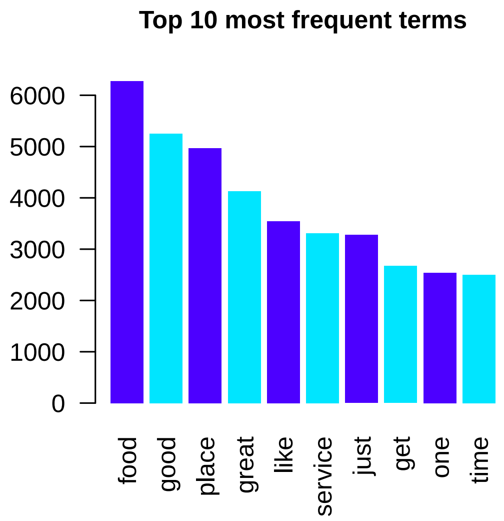
Methods and Data

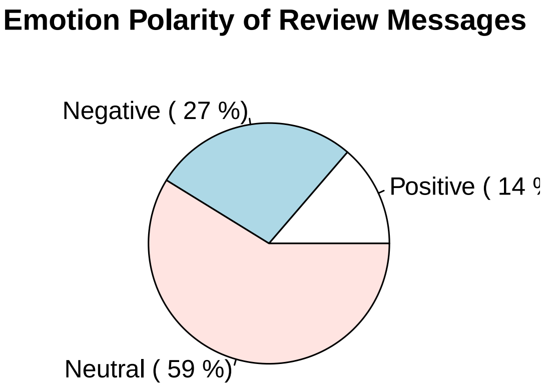
The dataset is obtained from the YELP website (http://www.yelp.com/dataset_challenge) and extracted. The format for the dataset is in `JSON` . JSON need special techniques to parse and read from it. Apache Hive is the best component which is capable read this format . Since the dataset required to have 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 . All the reviews and tips messages is filtered by removing the `numbers`,`symbols` and `unnecessary spaces` using R.

For the basic analysis , this evaluation requires a fair amount time to know about the dataset by performing 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 ?
- What is the impact to the ratings by `negative` , `positive` or `neutral` polarity ?

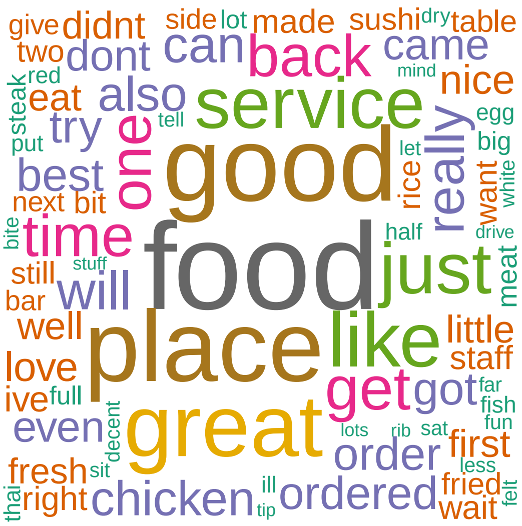
Below is the answers for the questions above . Top common words inside the review is regarding the `good food , good places` and also a `good services`. Reviewers whom visits seems very happy about the quality of food , services and places restaurant in Las Vegas. Most of the comments seems positively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





- Food Selection & Emotion Discovery

To discover what types of food have been reviewed most by the reviewer, the `word-cloud` plot is used to split the word frequencies. Since having 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 cane asily identify manually pick the food that we recognize as follows :-



From the above information we can make the assumption that the result shows the common food that always mention by the reviewer are `chicken` , `sushi` , `burger` , `pizza` , `tuna` , `salad` , `rice` , `shrimp` , etc. We use this list as the base of common food can be relate with the emotion of the reviewer. List of emotions that can be identified, such as `love`, `good`, `best`, `great`, `nice`, etc. Seems like this kind of food appear many times and the Last Vegas visitors or reviewer pretty happy with it. Some class type of food origin found in the list such as `Italian`, `Japanese` and `American` , etc. cuisine.

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

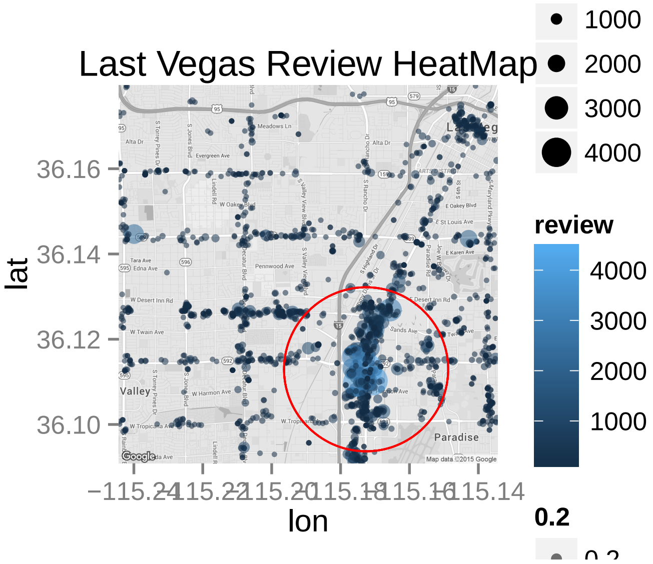
$$Score = \sum_{i=1}^{n} NumbersOfPositiveWords - \sum_{i=1}^{n} NumbersOfNegativeWords$$

- 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 to the `positive` and `negative` words is referred 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 :-



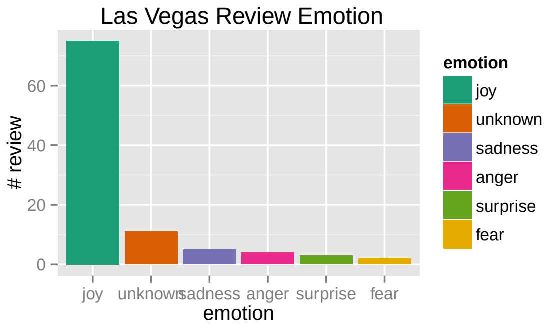
The red area is the central area of the review messages is focusing on. Major comments all over the year since 2004 is focused in that area. The location seems very strategic and it is only 2.5 miles from the `McCarran International Airport`. Most people will stay in this area and enjoyed the food before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

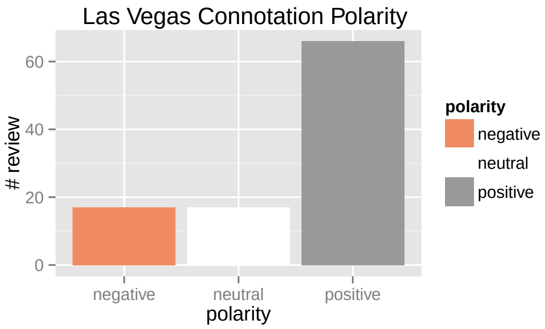
From the sample of the Las Vegas dataset , we found that more than 70% are happy `(joy)` emotion with the services, smaller portion are unhappy `(sadness, anger, fear)` emotion. The `(unknown)` emotion still needs deeper analysis to increase accuracy of the emotion classification.This smaller sample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

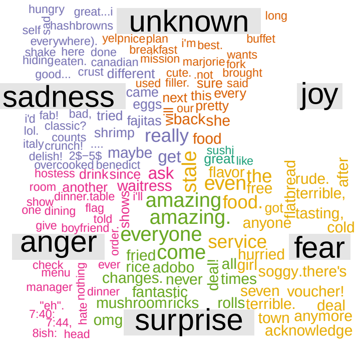
A `Naive Bayes` algorithm classifies each review message into subjectivity and polarity from the input. A subjectivity indicates the words are strongly or weakly subjective and the polarity indicates the weather the words are expressing a positive or negative sentiment. The illustration & table below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

Table 4: Sample of Emotion Classification Table

ANGER	DISGUST	FEAR	JOY	SADNESS	BEST_FIT
1.47	3.09	2.07	7.34	1.73	joy
1.47	3.09	2.07	1.03	7.34	sadness
1.47	3.09	2.07	7.34	1.73	joy
1.47	3.09	2.07	13.7	1.73	joy
7.34	3.09	2.07	7.34	13	sadness
1.47	3.09	2.07	1.03	1.73	NA







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 analysis techniques can be improved by three suggestions below :-

1) How `BigData` can help the analysis as a robust tool to support the high volume dataset for better and faster 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 , which believe can improve the model accuracy and performance.

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)