The Sentimen Evaluation of Customer Review in Las Vegas City Restaurants

Pozy Pak Ya

October 22, 2015

Abstract

From recent trends, many online reviews include a numerical or star rating that quantifies the satisfaction of the reviewer's experience. However, an objective mapping of this quantitative rating to the reviewer's textual description does not yet exist. In this paper, we explore models ranging from support vector machines to learning word vectors that capture the sentiment information of individual words in relation to ratings of an entire document. We use Yelp reviews for restaurants near particular universities to predict corresponding star ratings per review.

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.

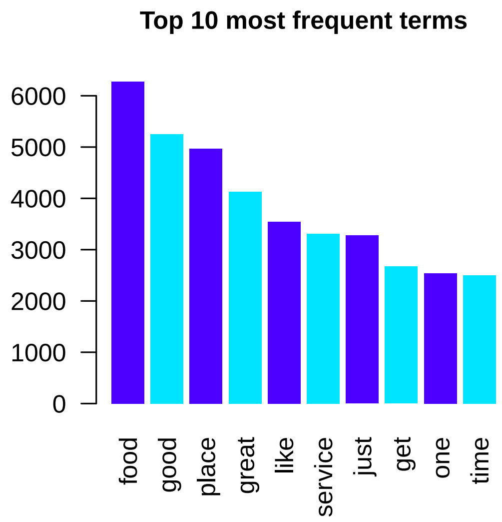
Methods and Data

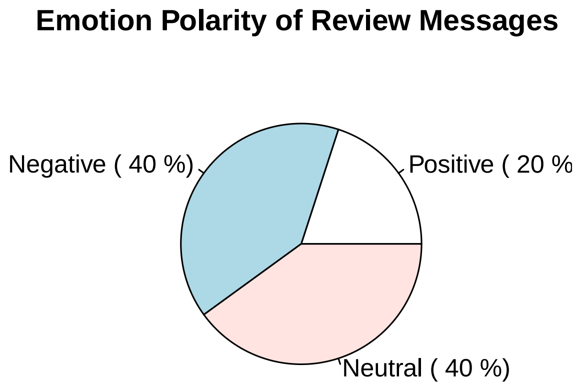
The dataset is obtained from the YELP website (http://www.yelp.com/dataset_challenge) and extracted. The format for the dataset is in `JSON` . JSON need special techniques to parse and read from it. Apache Hive is the best component which is capable read this format . Since the dataset required to have 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 conjuction 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.

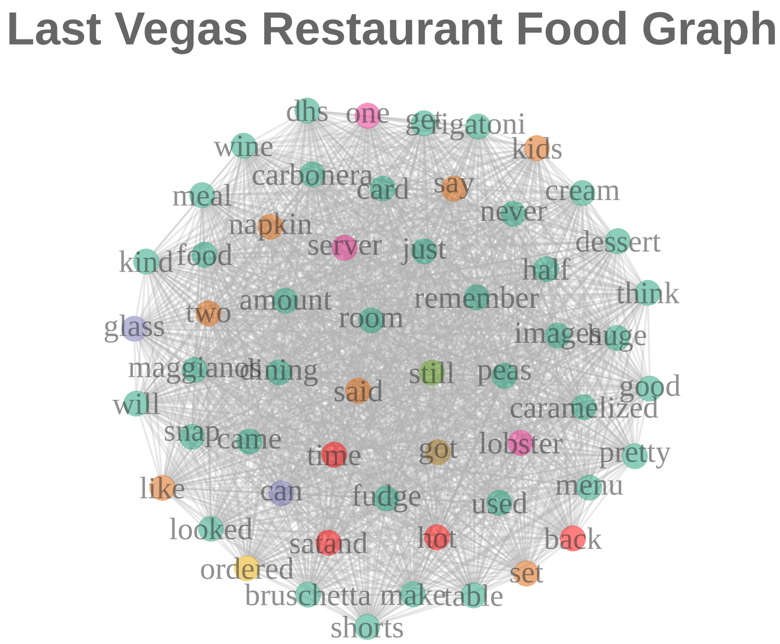




- Food Selection Discovery

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 :-

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` , `cheese` , `salad` , `rice` and `sauce`. We use this list as the base of common food can be relate with the emotion of the reviewer. List of emotion that can be identified , such as `friendly` , `best` , `well` , and `nice` . So, we entrusted with the food type listed and we want to have some idea what are their possible relationship . To achieve this , is to use the `word-graph` techniques for 8 groups of food and the result as follows :-



The graph shows that there is a few groups of words which their possible relationship and have the idea of the main term used.

Other interesting findings in this evaluation is to classify the reviewers ratings and the tips provided. The idea is to calculate the sentimen 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 :-

```
Score  =  Number of positive words  -  Number of negative words`
```

- 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 all analysis we can summarized by plotting the dispersion of the message size inside the map of Las Vegas restaurant. We the conclusion , we find out 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 :-

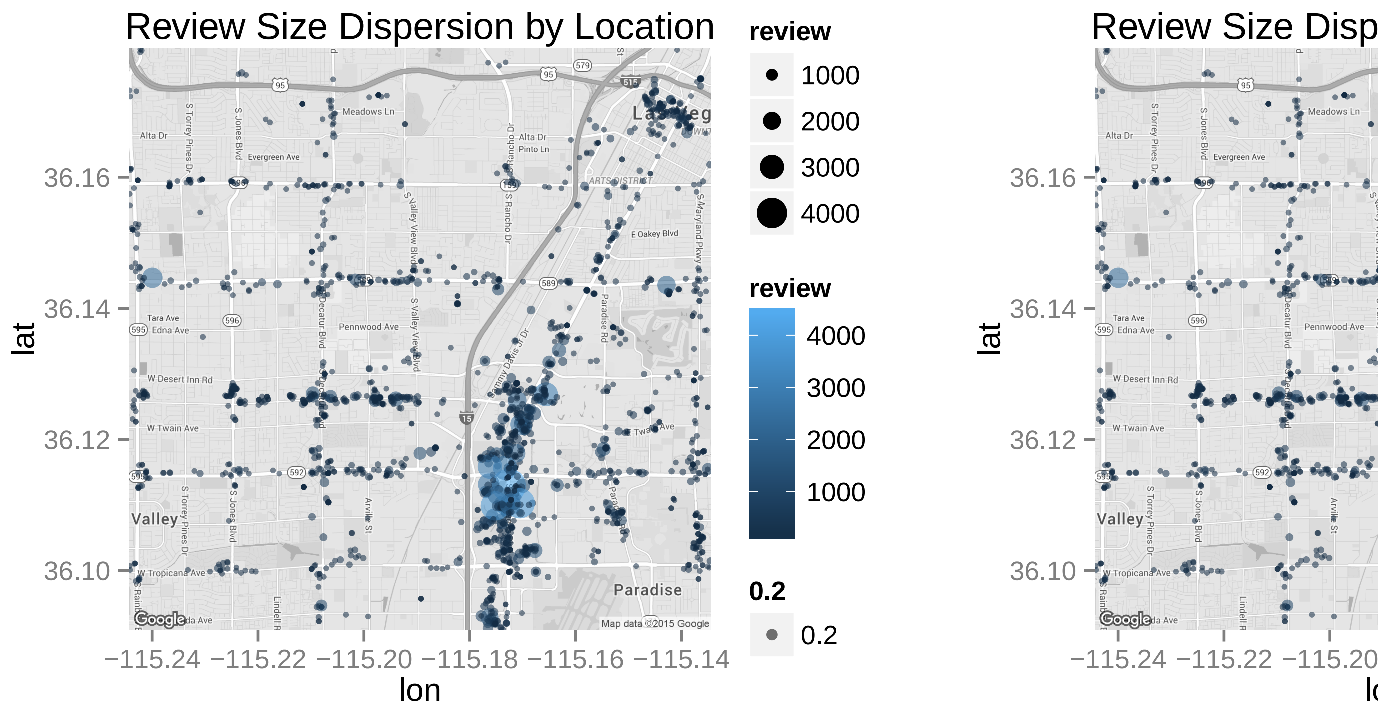


Figure 1 : Venue with greater than 390 review messages

Bagitau list of restaurant yang banyak impact dan types of food yang ada sesama mereka kalau ada relationship

The population is because the location is very strategic and nearest to the airport

Figure 2 : Venue with greater than 390 review messages

- Sentiment Analysis

Positive correlated with the positive emotion words

Figure 3 : Venue with weighted by the numbers of keywords

Sentiment analysis on keyword left 70% of the dataset with neutral sentiment. Sentiment was found to be 74% positive in nature which corresponds to about 22% of the total sample In contrast,26% was negative in nature which corresponds to only 8% of the total sample

Figure 4 : Venue with weighted by the numbers of keywords , with positive weight Figure 5 : Venue with weighted by the numbers of keywords , with negative weight Figure 6 : Total sentiment classification with positive , negative and neutral Figure 7 : Cloudwords postive and negative

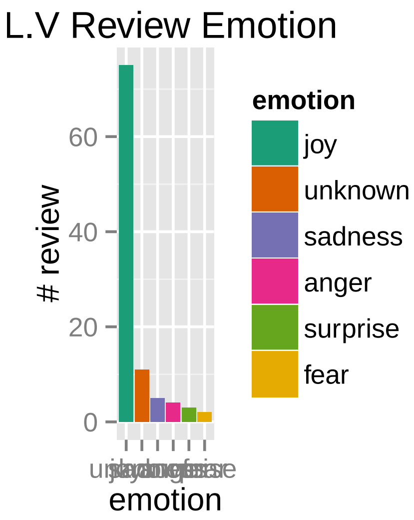
- Temporal Analysis

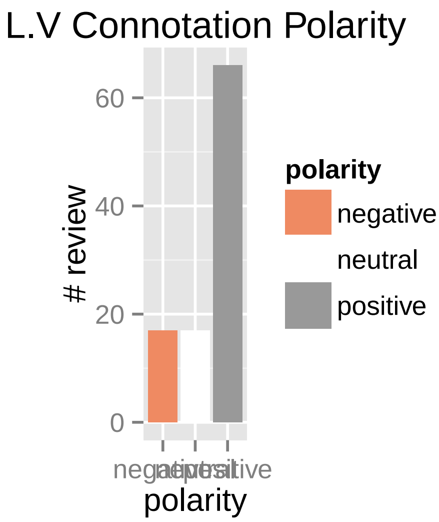
Figure 8 : Duration - Negative vs Positive + Neutral Figure 9 : Comparison by month

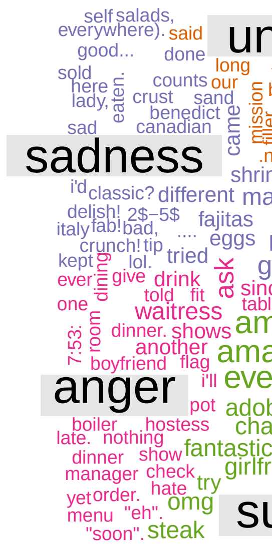
4. Method Used - Classification , Bayes

Is there any relation between the business type ?

The `HEAD` records for business types , average ratings and the average review count as follows :

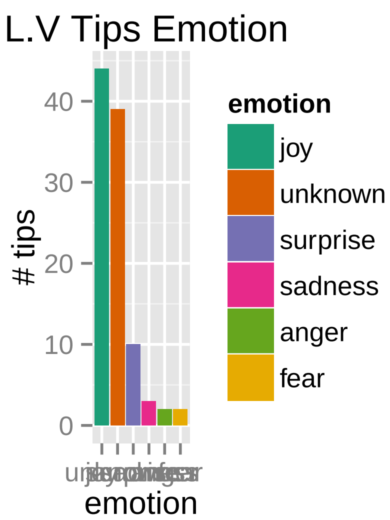


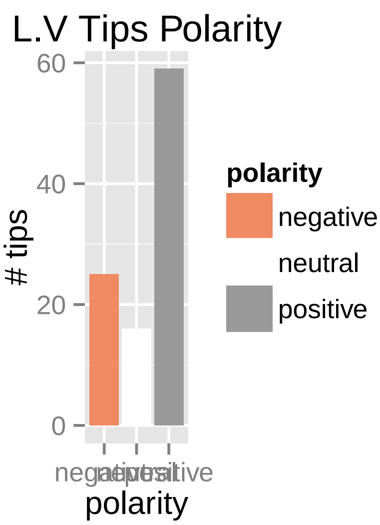


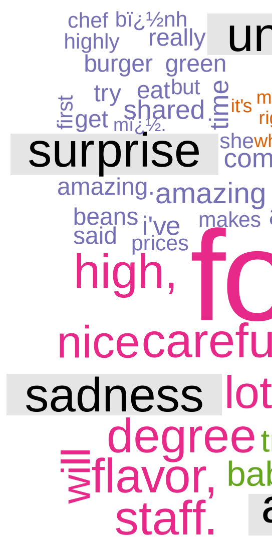


Tips

Cerita techniques yang dipakai di sini







Discussion

DISCUSSION The issue of using such data in a real planning situation raises several significant questions: 1) how to determine the extent of limited demographic information, data frequency, and the privacy concerns of Twitter users, 2) accessing and processing increasingly larger data, and 3) assuring a sample's diversity. In response, we believe that the three analysis techniques described above suggests that: 1) Big Data can become a robust tool, 2) Density can be coupled with temporal patterns to reveal dynamic sentiment monitored city wide offering detailed views of social trends. For issues related to better informed decision making, characterizing urban areas is essential. Through the union of data and location as exemplified in this study, multiple layers of realtime information can be displayed geographically, providing enhanced situational contexts and the deciphering of evolving social narratives. Odendaal (2006) acknowledges this premise, suggesting a growing movement in urban planning to utilize city narratives in the process of understanding place. The paper suggests that public life within digital social networks can tell an exceptionally well documented story.

In our experiment, we mapped the star ratings down to simplified 1 and 0 values, to signify a sharp polarity between positive and negative reviews. Initially, we had hoped to work towards a model that allowed us to make an entirely quantitative star rating measure of a review. We can design our model to better capture this information. We can also include more data by factoring in the three ratings per review provided by Yelp.

We have presented 3 algorithms chosen for their simplicity of implementation and run time efficiency. The results suggest that our classification-based approach performs better than numeric or ordinal regression approaches. Our next step is to verify these results with the more advanced algorithms outlined below.1. For many numeric regression problems, (boosted) classification trees have shown good performance.2. Several multi-threshold implementations of Support Vector Ordinal Regression are compared in Chu and Keerthi (2005). While they are more principled than the Perceptron-based PRank, their implementation is significantly more complex. A simpler approach that performs regression using a single classifier extracts extended examples from the original examples (Li and Lin, 2007). 3. Among classification-based approaches, nested binary classifiers have been proposed (Frank and Hall, 2001) to take into account the ordering information, but the prediction procedure based on classifier score difference is ad-hoc.

Textual reviews for different products and services are abundant. Still, when trying to make a buy decision, getting sufficient and reliable information can be a daunting task. In this work, instead of a single overall rating we focus on providing ratings for multiple aspects of the product/service. Since most textual reviews are rarely accompanied by multiple aspect ratings, such ratings must be deduced from predictive models. Several authors in the past have studied this problem using both classification and regression models. In this work we show that even though the aspect rating problem seems like a regression problem, maximum entropy classification models perform the best. Results also show a strong inter-dependence in the way users rate different aspects.

https://sites.google.com/site/miningtwitter/questions/talking-about/given-topic