Yelp Ratings System on Steroids [title is a work in progress :/]

Ammar Karim, XiuXiang Jin, Zeqiang Lin, YuHao Liu

Csc 59866 Fall 2017

Senior Project Design Report

I. Introduction

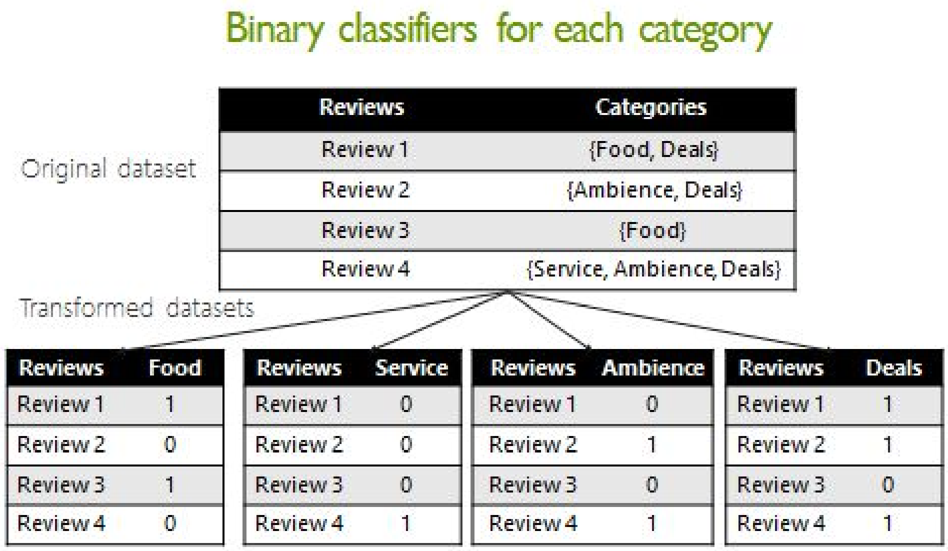
When it comes to choosing a restaurant, Yelp is the first destination foodies go when looking for reviews and ratings. Together, these ratings and reviews allow potential customers to decide where to go. Yelp provides a single 5 star rating. However, it is difficult to understand the context in which a certain rating was given. Reviews can help clarify ratings but poring through multiple reviews to understand the context is not feasible each time someone wants to go out to eat.

When someone is looking through reviews and ratings, he or she already has personal preferences with regards to what is important in a restaurant. Some people value the decor and friendliness of staff highly while others are simply interested in the quality of the food.

We seek to create an application that will extract the categories that a user cares about. Then using these categories we will analyze reviews of a restaurant and extract a rating for each of the categories that the user cares about. We can further extend this system by finding restaurants that a user may care about and providing a recommendation based on the improved and categorized rating we extracted.

II. Background

Categorization of overall reviews for a restaurant into relevant parts involves using support vector machine to extract sentiments from each review. The Mondego group from Donald Bren School of Information and Computer Sciences have built a Data Mining Project on Yelp Challenge. Their approach was to classify a sentence of yelp review rating into numerous categories for a restaurant. To start out, their team first obtained 10,000 restaurant reviews from the yelp released dataset and divided them into 5 bins. They assigned 6 researcher to manually annotated those reviews and come up with the 5 types of impression for a restaurant: food, ambience, service, deals, and worthiness. 80% of those data are used as training and the other 20% for testing. Then, two types of feature are extracted from the review: the first one is the user rating with 1 to 2 stars indicating bad, 3 stars for moderate, and 4 to 5 stars for good. The second feature consist unigram, bigram and trigrams with frequency greater than 300 using arff files of machine learning project. Because a review might mention multiple impression for a restaurant, multiple categories binary classifier is needed to get the final prediction. Figure 1 gives an example of how 4 reviews containing multiple categories will be divided. After that, Support Vector Machine can be used to independently classify the subset of categories and union them together with voting and come up with the result as shown in figure 2.



Reviews	Categories
Review 1	{Food, Deals}
Review 2	{Ambience, Deals}
Review 3	{Food}
Review 4	{Service, Ambience, Deals}

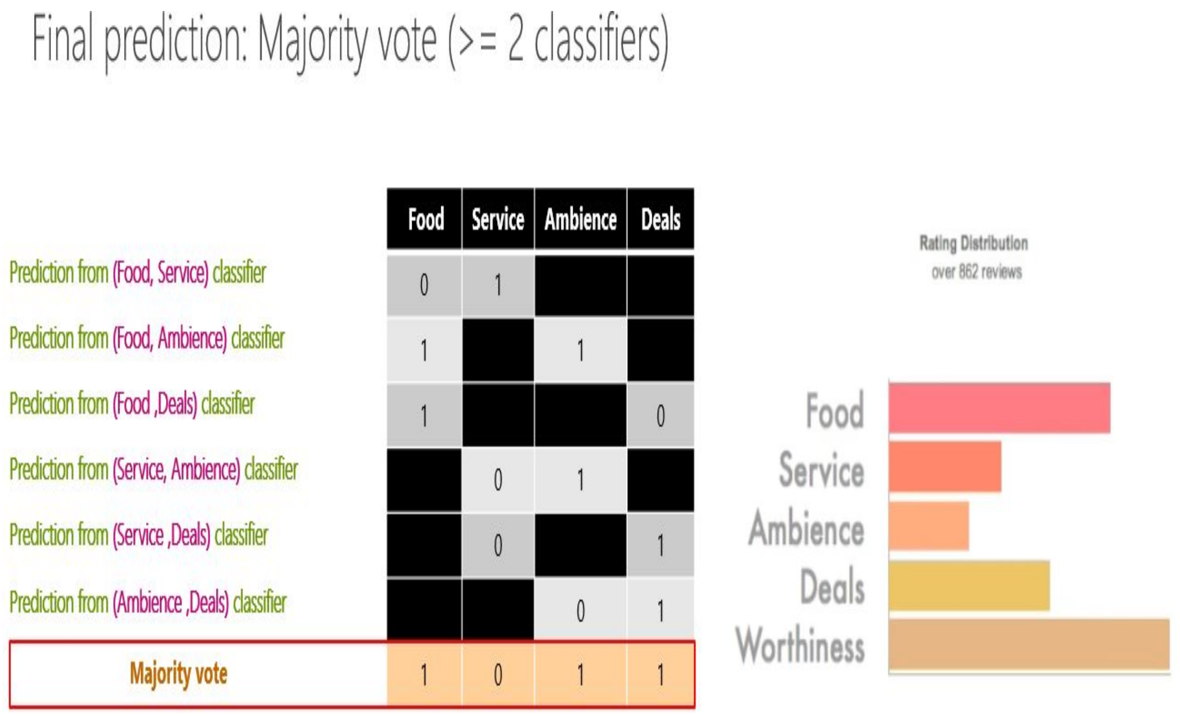
Reviews	Food
Review 1	1
Review 2	0
Review 3	1
Review 4	0

Reviews	Service
Review 1	0
Review 2	0
Review 3	0
Review 4	1

Reviews	Ambience
Review 1	0
Review 2	1
Review 3	0
Review 4	1

Reviews	Deals
Review 1	1
Review 2	1
Review 3	0
Review 4	1

Figure 1



Final prediction: Majority vote (>= 2 classifiers)

	Food	Service	Ambience	Deals
Prediction from (Food, Service) classifier	0	1	■	■
Prediction from (Food, Ambience) classifier	1	■	1	■
Prediction from (Food ,Deals) classifier	1	■	■	0
Prediction from (Service, Ambience) classifier	■	0	1	■
Prediction from (Service ,Deals) classifier	■	0	■	1
Prediction from (Ambience ,Deals) classifier	■	■	0	1
Majority vote	1	0	1	1

Figure 2.

III. Team Organization

Our team were organized with the responsibility stated as followed: XiuXiang will be doing the background research on what others have accomplished on their project for improving the yelp review and share them with the group to get an inspiration out of it. He will also be analyzing the methodologies and features available out there that could be applied to our project. Yu Hao is focusing Natural Language Processing, he mainly focusing on keyword extraction. After that, he will work on machine learning algorithms. Ammar will focus on data collection and structuring to be used for NLP and machine learning algorithms. Zeqiang will work on the design of classifier which for the final prediction.

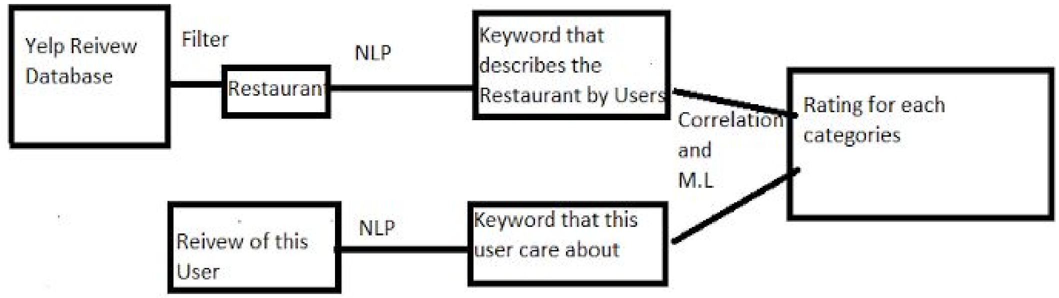
III. I. Data Collection

We plan to have two data sources: initially we will collect all the reviews a user has made and all the reviews made about a certain restaurant. Yelp does have an api that provides ratings and user information but does not provide all the reviews of a restaurant. We have decided to create a web scraper that will collect the reviews but this is not an ideal solution. A similar scraper will need to be made to get the user reviews.

III. II. Structure and Algorithm

The Natural Language Processing algorithm that we are using is Key Extraction. It is one of the most common algorithms in Natural Language Processing. The three main components of Key extraction algorithm are candidate selection, properties calculation, and scoring and selection keyword. The candidate selection is a processing to select candidate that is possible to be keyword. Therefore, it goes through all the words, phrase, and terms. Then for every candidate, it does a properties calculation. This means, it selects keywords based on the properties of each candidate. For example, if a words, phrase, and terms that appears in title, therefore, it is important and it will be a keyword. The lastly, it scores and selects keywords. There are two ways to do it. One way is using some type of formula to generation the value of each keywords. For example, there is a library in python called RAKE, is doing key extraction algorithm using some type of formula. Another way is using machine algorithm. For example, there is a library in Java called Maui used a machine learning toolkit called WEKA to do keyword extraction. A example of keyword extraction algorithm: when a review is “Good sport.. Nice ambience.. nice decor.. Kind of pricey. Good thing I wasn’t one of those that got to wait in a long line”. Therefore, based on this statement, the potential keywords will be spot, ambience, decor, pricey, and wait.

IV. System Prototype


Yelp Reivew Database
Filter
Restauran
NLP
Keyword that describes the Restaurant by Users
Correlation and M.L
Rating for each categories
Reivew of this User
NLP
Keyword that this user care about

This is our prototype structure of our project. First we get the reviews from the Yelp review databases on specific restaurant. Then we will apply Natural Language Processing, which is key extraction algorithm, to receive the important keywords that all users informed about this restaurant. On other hand, we will do the same stuff to the user. Ideally, we will use the Yelp user review database, but for convenience, we will mock some user reviews for testing. We also apply NLP to user reviews and we get the keywords that this user care about the restaurant. Then we combine those two groups of keyword, and do some machine learning algorithms and apply some classifiers to generate a rating of each categories that user care about this restaurant.

V. Conclusion

This system is mainly focus on using NLP and machine learning to reinforce the functionality of Yelp reviews. Basically Yelp reviews are not categories into different types now, when users read those reviews, especially long reviews, they may not get the information they want. By separating reviews into different categories and assign them scores, users can understand the advantage and disadvantage of that restaurant based on scores. It saves users' time and help them in reviewing comments. By now the system designed to divide reviews into five categories, and it can be divided into more categories such as different type of meals to enforce the functionality of the system.

References

Medelyan, Alyona. NLP keyword extraction tutorial with RAKE and Maui.
https://www.airpair.com/nlp/keyword-extraction-tutorial

Sajnani, Hitesh, et al. "Yelp Dataset Challenge." The Yelp Dataset Challenge - Multilabel Classification of Yelp Reviews into Relevant Categories, www.ics.uci.edu/~vpsaini/.

"The Best of Yelp." Yelp, www.yelp.com/c/manhattan/restaurants.
https://www.yelp.com/c/manhattan/restaurants