Buon Appetito : Recommending Personalized Menus with Sentiment Analysis and Graph Discovery

MICHELE TREVISIOL , Yahoo Labs
LUCA CHIARANDINI , Google
RICARDO BAEZA-YATES , Yahoo Labs

This is the abstract.

Additional Key Words and Phrases: Recommender System, Menu Builder, User Emotion

ACM Reference Format:
Michele Trevisiol, Luca Chiarandini and Ricardo Baeza-Yates, 2015. Buon Appetito .... *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 39 (March 2010), 0 pages.
DOI: 0000001.0000001

1. INTRODUCTION

✍ **TO-DO:**
- ✗ paper goal, why it is useful
- ✗ novelty of the idea
- ✗ approach: graph, sentiments, etc.
- ✗ results

2. RELATED WORK

✍ **TO-DO:**
- ✗ same as before
- ✗ add more related to graph
- ✗ add our previous paper

3. ANALYSIS

✍ **TO-DO:** pasted the old old section

This research is partially supported by the Ministry of Science and Innovation of Spain.

DOI: 0000001.0000001

✍ **TO-DO:**

✗ explain + overview of the dataset (no. business, no. users, no. reviews, cities, size, . . .)

✗ add stats for each reviews (avg. length, avg. no. sentiments, avg. no. food items, . . .)

✗ overview of sentiments (LIWC dictionary vs. TextBlob, review- vs. sentence-level, . . .)

✗ food recognition (dishes vs. ingredetients, explain graph, advantages, performance, . . .)

✗ taste changes over time (user/business profile, check variation over time, over cities, over cultures, . . .)

In this Section we describe how we built our dataset and present some interesting results aimed at validating the soundness of the process. First of all, we will describe the dimensions along which we processed the Yelp reviews: we are interested in the sentiment, the social environment and the food. Secondly, we will analyze each dimension separately and, finally, we will combine all dimensions, highlighting interesting relationships between them.

3.1. Dimensions of the Analysis

Venue reviews could be analyzed under different points of view, since many interesting aspects could be extracted from the text of the reviews.

Since we are interested in recommending personalized menus for restaurants, *food* is certainly a dimension we would like to explore. We aim at detecting popular foods, understanding if and how food varies in terms of category of restaurant (*e.g.* Mexican, Japanese, Italian, *etc.*), and extracting frequent combination of foods, thus looking for menus.

Food alone is not enough. Reviews have a *sentiment*, that is, they could be positive reviews or negative reviews. Sentiment is therefore an important dimension to look into, especially related to the rating of the reviews (*i.e.*, stars that the person assigned to the place).

Finally, even if not essential as the previous dimensions, *social environment*, *i.e.* the people involved in the review, is an interesting dimension for the analysis. Indeed, the social context influences the behavior of people during a meal. For example, people may order different menus or have different expectations when dining with the family, with friends or with colleagues. Therefore, we expect that the content of the review may depend on this factor.

To summarize, the dimensions we will consider are:

— *Sentiment* S: amount in which the text is conceiving a positive or negative judgement;
— *Food* F: set of dishes or food present in the review;
— *Social Environment* E: relationships between people, measured by the degree of closeness (*e.g.* a partner is closer, whereas a colleague is distant).

We will now analyze each dimension separately by highlighting how we extracted it from the text and showing interesting findings for each one of them.

3.2. Preprocessing

Since reviews are written in natural language, we perform a preprocessing aimed at reducing noise and sparsity. First of all, we split a review in sentences. For each senten-

$S = -1$	$-1 < S < +1$	$S = +1$

Table I: Comparison of sentiment S in review- and sentence-level aggregation.

ce, we remove stopwords such as prepositions (*e.g.*, "to", "for", *etc.*), conjunctions (*e.g.*, "and", "or", *etc.*), pronouns and other common words. We then lemmatize all words in the sentence in order to remove the number of words and therefore the sparsity.

✍ **TO-DO:** Michele: check

3.3. Sentiment S

Sentiment analysis is the use of Natural Language Processing to identify subjective information from text. In this work, we are interested in understanding the *polarity* of a text, *i.e.* the amount in which it is positive or negative.

There are many method to do sentiment analysis. We will adopt a simple mechanism that relies in recognizing *polar words*, *i.e.* words that convey a positive or negative emotion, and use them to score the text. We use LIWC 2007 [**?**] dictionary of sentimentally-annotated words. Each word has a number of facets connected to it (*e.g.*, grammatical features, topics, *etc.*). Among all facets, we focus on those about polarity.

Given a text, we are able to score the positiveness by counting the occurrence of polarity words in it. The sentiment score of a text is given by Equation **??**:

$$S = \frac{p - n}{p + n} \tag{1}$$

where p is the number of positive words, and n is the number of negative words.

3.3.1. Sentence- vs. Review-level Sentiment Analysis. The first question we ask is whether dividing the text of the reviews in sentences affects the results. To answer this, we compare the distribution of sentiment in the whole review and the one of its sentences.

Table **??** shows the comparison between the two cases. We show three cases: (1) *negative*, when $S = -1$; (2) *mixed*, when $0 < S < 1$; and (3) *positive*, when $S = 1$. We can see that, in the case of sentiment of sentences, there is a majority of positive and negative, while review sentiments are more mixed. Indeed, mixed sentiments occur only in 12% of the sentences, against 43% in the case of review-level aggregation.

We conclude that splitting by sentence allows us to get a more precise, clean, and localized characterization of the text.

3.3.2. Sentiment and Ratings. It is natural to expect that the sentiment of a review is related to the ratings given by users. Figure **??** shows the distribution of sentiments for each rating. We can see that the amount of positive reviews decreases and that the amount of negative reviews decreases with the rating. The tendency is towards positive reviews, which is a very well known effect called *Pollyanna principle* [**?**].

A small amount of purely negative reviews ($S = -1$) are always present, even for highly-rated reviews. Observing such reviews manually, we understand that they are not related to the place itself, but rather to contingencies. For example, the following reviewer is disappointed that a cafe is closed, but gave the maximum review anyway:

> [...] I stopped by two days ago unaware that they had closed. I am severely bummed. This place is irreplaceable! [...]

3.4. Food F

Food, along with the quality of service, is the most important aspect of restaurant and cafe reviews. People write about food in their reviews and often share their favorite

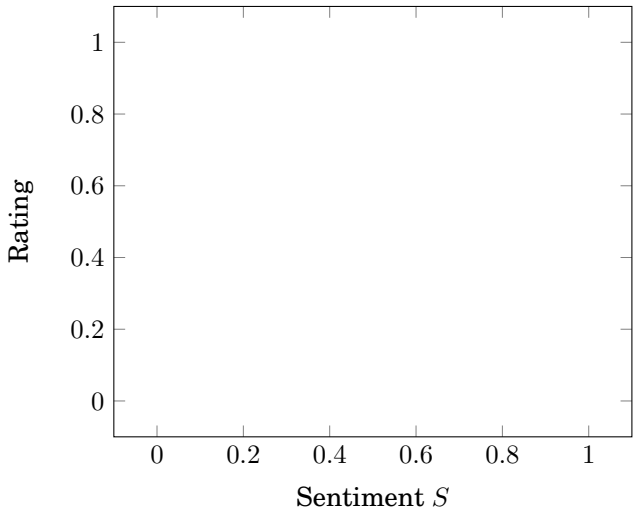

Fig. 1: Comparison between sentiment S and rating R (expressed in number of stars, 1 is worse, 5 is best). The area of the circle represents the amount of reviews.

Food

Table II: Most frequent foods in the dataset.

menu at a particular place. Extracting food from text is a difficult task due to the large amount of ingredients and local expressions.

We implemented a basic method which captures the most popular foods in reviews based on a dictionary extracted from three publicly available sources:

— *Oregon State University Food Glossary*[1]: this is a multi-language glossary of food which contains ingredients as well as scientific names. Using a web-crawler, we built a dictionary based on the titles of the food pages in the glossary.
— *WordNet*[2]: WordNet is a large lexical database of English. We built a dictionary containing all nouns in the "food" group.
— *BBC Food*[3]: BBC Food is a web portal of recipes and ingredients. It contains a large amount of recipes written in English. We crawled the pages and extracted all ingredients and recipes. For each page, we crawled the displayed image and the description. Also, for each recipe, we were able to crawl all its ingredients. The final dictionary consists of around 9000 items.

After building the dictionary, we lemmatize all words, we manually remove some noise and we find such words in the text of the reviews.

3.4.1. Statistics. Table **??** shows the most frequent food words in the dataset, alongside with the percentage of occurrences. A breakdown by type of restaurant is provided in Table **??**. We can see that the food which appear in the list are indeed typical of the particular cuisine.

[1]http://food.oregonstate.edu/
[2]http://wordnet.princeton.edu/
[3]http://www.bbc.co.uk/food/

American		Italian	
Mexican		Chinese	

Table III: Most frequent foods for various type of restaurants.

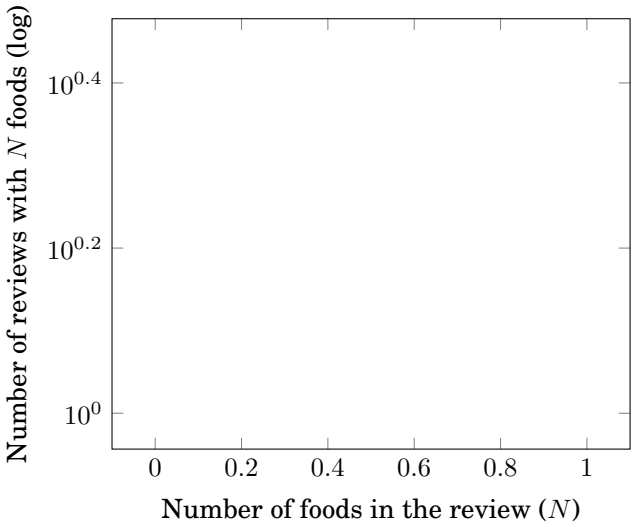

Fig. 2: Number of food per reviews.

As for the coverage of the foods in the reviews, on average, we detect 4.7 food words per review. Figure **??** shows the head of the distribution of food per review. The number of food per review is uniform across ratings and types of restaurants.

3.5. Social Environment E

✍ **TO-DO**: How to extract the dimensions from reviews

✍ **TO-DO**: Show validity of feature by plotting distributions. Show examples

3.6. Food and Sentiment

Having inspected every dimension by itself, it is now time to analyze dimensions jointly. We start by analyzing maybe the most interesting one for our goal: sentiment and food.

3.6.1. Sentence-level Food Sentiment. We observed in Section **??** that extracting sentiment from sentences gives a more localized and clear signal than extracting it from the whole review text. In addition, being able to detect the sentiment of particular sentences in the reviews allows us to better connect the sentiment to the food words. It is indeed quite common (57% of reviews, see Table **??**) for people to write a mixed-sentiment reviews. This often happens when reviewing more than one dish, as for example:

> [...] Pizza crust & toppings are excellent. However the pizza sauce was too salty. [...]

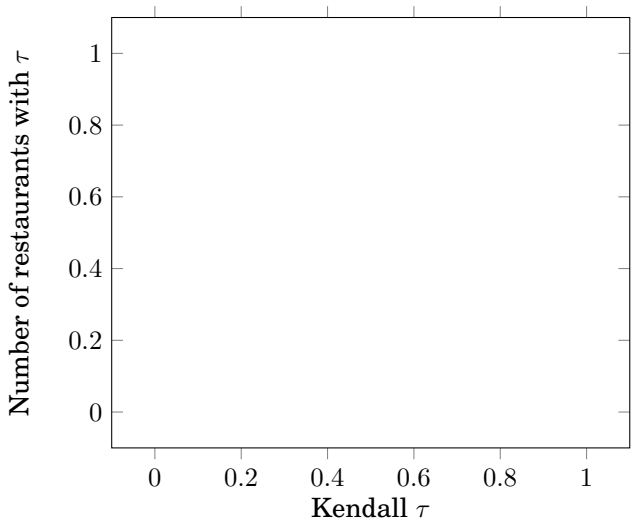

Fig. 3: Distribution of Kendall τ.

Bearing this in mind, we assign to each occurrence of a food word the sentiment S of the sentence it belongs.

In order to evaluate the way in which we assign sentiment to food words, we compare it to the ratings given by users. We build the ranking of food words based solely on the sentence-level sentiments r_S, and we compare it to the ranking we would obtain using the ratings of the reviews r_R.

First of all, we score each food by summing up the contributions of its occurrences: for r_S, each sentence containing the food contributes with its sentiment, for r_R each review containing the food contributes with its rating. Given the score of the food, we compute the two rankings r_S and r_R. Since the preference of food may depend on the particular restaurant, we build r_S and r_R for each restaurant that has been reviewed at least $N = 20$ times.

We then compare the rankings using Kendall rank correlation coefficient τ [**?**]. Figure **??** shows the distribution of the value of the coefficient of the restaurants. We can see that the two rankings show a slight agreement, with an average tau of

✍ **TO-DO**: PUT VALUE

.

3.7. From Dishes to Menus

In the previous section we built a ranking of dishes for each restaurant based on sentence-level sentiment analysis. A ranking of food could be already useful by itself, since it could be used to recommend to customers the best dishes for a particular place. However, many restaurants provide a menu, not just individual dishes. A good menu is not only composed by good dishes, but relies in the harmonic combination of flavors, It is therefore natural to expand our analysis towards menus.

A menu is a sequence of dishes which are served during a meal. Many cuisine specify an ordering of dishes. However, for simplicity, we will discard such orderings, although it may be possible to infer them from the reviews. In this work, a menu is simply a set of dishes. We are interested in detecting *good* menus. The meaning of the word “good”

can be multiple. We will start by extracting *frequent* menus, *i.e.* menus that people often choose, and expand our algorithms to consider the quality of menus, intended as people's opinion.

3.7.1. Extracting Frequent Menus. Frequent Itemset Mining is the task of extracting sets of items which occur frequently in a database. The Apriori algorithm [**?**] is one of the most known algorithm to perform this task. It receives as input a set of sets T , which we call transactions, and a number,$minSup \in [0, 1]$, and returns the sets which are subset of at least $minSup \cdot |T|$ elements of T. Apriori appears as the natural choice for our problem of detecting frequent menus.

The set of transactions T contains one set for each review, containing the food words that appear in it. We ran Apriori with $minSup = 0.05$ and extract the most popular item sets for each restaurant category. The results are displayed in the first column of Table

✍ **TO-DO**: ref

3.7.2. Extracting Frequent and Positive Menus. Until this point, we completely discarded a very useful resource: people's opinion. We are indeed interested in extracting not only menus that occur frequently, but also that people like.

We use an extension of the Apriori algorithm [**?**] that deals with the case in which the transactions in T are *fuzzy*. Fuzzy sets are sets whose elements have degrees of membership. The algorithm is designed to find those itemsets that are frequent in the sense of fuzzy sets, *i.e.* those that are strong members of many transactions.

This algorithm is suitable for our case if we interpret set membership as preference. As in the case of frequent menus, transactions are reviews and they contain the food words that appear in them. What we add in this case is the fact that food words belong to the transaction depending on their sentiment. For each review, the membership of each food word is the averaged sentiment that it receives in the sentences, normalized to fall in the interval $[0, 1]$. Different food words in the same transaction may have different degrees of membership. If a food has a negative sentiment, it will have membership of 0. On the contrary, if it has fully positive sentiment, it will have membership of 1. Note that full negative sentiment is equivalent to the food never occurring in the review. Discarding food that people dislike does not change our approach since we are interested in the menu that people liked most.

Having built the fuzzy transactions from the reviews, we run the fuzzy version of Apriori, with again $minSup = 0.05$. The results are displayed in the second column of Table

✍ **TO-DO**: ref

In order to compare the two versions of Apriori, we rank the item sets by their support and compare the two rankings using Kendall rank correlation coefficient. The results are show in Figure **??**.

✍ **TO-DO**: Explain

3.8. All together

✍ **TO-DO**: Put all features in a single matrix. Is the matrix sparse?

✍ **TO-DO:** Analysis and statistics on the dataset:

— Relation with reviews, starts
— About Check-in
— What about the Sex guessed by the name of the user? We should ask Eduardo about the DB he knows about this.

☞ **NOTE:** If we do not use the gender, this is not really useful

4. EXPERIMENTS

✍ **TO-DO:**

- ✗ recommendation approaches:
  - ✗ popularity
  - ✗ serendipity
  - ✗ mix (business−profile, user−profile)
  - ✗ CF
  - ✗ by cities/cultures if analysis makes sense
- ✗ matching techniques:
  - ✗ using only dishes
  - ✗ using dishes + ingredients
  - ✗ taking a subset of the graph
  - ✗ word2vec
  - ✗ tf−idf/bm25
- ✗ evaluation metrics:
  - ✗ F−score
  - ✗ Precision
  - ✗ nDCG (if ranking matters)

Which is the ground truth ? Since we have to perform different experiments we need to discuss this very carefully..

✍ **TO-DO:** which is the best evaluation metric? precision? recall? f-1? and why..

Emotions Prediction

The goal here is to prove how good are the emotions. They should behave in line with the starts. However we can also find some ranges of no. of posemo and no. of starts. However this Section motivate us to go to the next one.

— Which is the predictive power of the emotions? How do they behave compared to the stars?
— Experiments (train/test set) on the efficiency of our emotions recognition approach

Menu Prediction

The goal here is to predict what the user is going to eat, or in other word, which will be his menu. This is very challenging and it seems a very innovative contribution. Anyway in this case we use the sentence-sentiments to have a rate for each food the user ate in the past. Since we proved before that the sentiments are working similarly as the starts, we use them in a more fine-grade since we can extract the sentiment

(*i.e.*, rate) of each sentence.

Process:

— Build a user menu profile (with the things he ate in the paste - no ratings!).
— For each restaurant build a rated menu (by sentence sentiments), where for each food we have the frequency and the avg-rate.
— Split the dataset into train and test set in a smart way: be sure the user in the test set made *enough* reviews in the train set and in both of the set they write something about food.
— Try to predict which things the user are going to consume in the test set.

5. CONCLUSION

Future work:

— *viceversa*: given menu items, recommend the most suitable place