Data-Driven Decisions:
Ranking of Dishes and Restaurants for Their Recommendation

1. Introduction

Below is a description of the principles that I used to develop my solutions. It is based on answers to the three questions listed below. Note that this brief report was prepared within a limited timeframe; it can be substantially improved if additional research is done.

Question 1. *Given a cuisine and a set of candidate dish names of the cuisine, how do we quantify the popularity of a dish? How can we discover the popular dishes that are liked by many reviewers? What kind of dishes should be ranked higher in general if we are to recommend dishes of a cuisine for people to try? Would the number of times a dish is mentioned in all the reviews be a better indicator of a popular dish than the number of restaurants whose reviews mentioned the dish?*

The key word here is “liked”. The sheer number of reviews is not capable of describing how much we like or dislike a dish or a restaurant. But this can be derived from the star ratings of reviews that mention a dish or are about a specific restaurant. Another intuitive metric could be a sentiment analysis of reviews that mention them. If both metrics coincide, this can be a positive indication of the meaningfulness of results obtained. I came up with my own ranking function which yielded some interesting results described in this report.

As for the number of times a dish is mentioned in all reviews vs. the number of restaurants whose reviews mention it, the more restaurants mention a dish, the more popular it is, but I would still analyze not the overall number of times a dish is mentioned, but the number of times it is mentioned positively (again, rank and sentiment) because a dish may be also mentioned when people dislike it.

Question 2. *For people who are interested in a particular dish or a certain type of dishes, which restaurants should be recommended? How can we design a ranking function based on the reviews of the restaurants that mention the particular dish(es)? Should a restaurant with more dish name occurrences be ranked higher than one with more unique dish names?*

Again, quality is more important than quantity here. The restaurants that should be recommended to those interested in a particular dish should be those that have better reviews for this dish. I would use the same ranking function as in Question 1, but apply it differently - instead of ranking dishes across all restaurants, I would rank all selected candidate dishes for a particular restaurant and then combine these dish ranks to get the rank of the restaurant with respect to the quality of the candidate dishes.

Question 3. *How can you visualize the recommended dishes for a cuisine and the recommended restaurants for particular dishes to make them as useful as possible to users? How can the visualization be incorporated into a usable system?*

The best way to visualize recommended dishes would be to sort them in the descending/ascending order and present to the user in the form of a list or a bar chart, maybe with some additional brief information which pops up when the user points the mouse at the dish - a so called tool tip at the mouseover event. If the users are interested, they can click on the dish and see more detailed information about it, maybe even a list of restaurants for which this dish is ranked the highest.

An additional future improvement of this system would be deriving the location information from the reviews file, for instance on the basis of the address or geographic coordinates (both are available in the Yelp reviews), and then recommending restaurants for a user-selected dish in the user-selected location.

2. Implementation

2.1 Dish List

I tried to make the dish list to be used by my system as all-inclusive as possible in order to investigate a wide gamut of Italian dishes without losing any potential information and to see what dishes are popular, as well as what restaurants are worth going to. I used the manually annotated file from Task 3 including all additional Italian dishes that I added to it through my own research plus the top 2000 most frequent entries from the Autophrase mining results (i.e. 90 to 100% probability that they are Italian dishes) and 800 entries from the Word2vec mining results because these two algorithms provided the best results during the Task 3 implementation. I had to manually annotate the dishes from the last two sources as they contained some extraneous non-dish related entries.

After that I sorted the list and removed duplicates using OpenRefine. If I had to decide which dish name variant to leave in the list, I did this for shorter dish names because if, for example, I search for a string "chicken parm", I will also find all occurrences of "chicken parmesan", "chicken parmigiano", etc., but not vice versa. For the same reason I preferred singular over plural because plural always includes singular (naturally, this is true only for regular nouns). In the end I had what appeared to be a comprehensive list of approximately 550 Italian dishes.

2.2 Ranking Dishes and Restaurants

As suggested in the Rubric, I made an extra effort to consider sentiment of comments by using both ratings and sentiment tagging of reviews (and to see if there is a correlation between them). I came up with my own ranking function which I used both for dishes and restaurants, but applied it differently. I didn't find this function in any online sources, but I did not make an extensive search because of limited time.

First of all, I was simultaneously collecting information about ratings and sentiment polarity for each combination of "restaurant + dish" by traversing all the reviews, looking for preselected candidate dishes in them, and noting what restaurant each review belongs to.

For sentiment analysis, I used the textblob module in Python, i.e. for each combination "restaurant + dish" I extracted the sentiment polarities of all the corresponding reviews and then calculated the mean sentiment polarity. The sentiment polarities were measured for the entire review text in order to establish the overall positive or negative context. As an alternative, it is also possible to measure the sentiment polarity for a specific sentence where the dish name is mentioned. But in my opinion, both methods have a certain subjectivity associated with them.

For ranking, the main idea was to use a balanced function in the range of [-1, 1] which is based on the star rating of each review in which a particular dish is mentioned to have been served in a particular restaurant. A completely negative rank is -1, and a completely positive rank is 1 with 0 being neutral. So, for each dish from my preselected candidate dish list from Section 2.1 I found all the reviews where it is mentioned and created a list in which each element is represented by the following data structure:

[***restaurant***,
dish,
[***list of all star ratings*** *from all reviews for "this restaurant + this dish"*],
[***list of all sentiment polarities*** *for all reviews for "this restaurant + this dish"*]
]

The combination "restaurant + dish" is unique and was used as the key. Then, I used information from the third element (list of star ratings) to calculate the average rank **for each pair "restaurant + dish"** using the following formulas:

negative rating = count of **1&2-star** ratings / count of all **1,2,4,5-star rating**

positive rating = count of **4&5-star** ratings / count of all **1,2,4,5-star rating**
(as you can see, I did not use 3-star ratings in the numerator as I consider them too neutral, sometimes even noise. Per Professor Zhai's suggestion, I also did not 3-star ratings in the denominator as they would make results less intuitive)

average rank = positive rating - negative rating

The division by the number of all review is necessary to normalize the ratings because now they do not depend on the number of reviews. Otherwise, the dishes/restaurants with more reviews would have had higher ratings. If all the reviews for this dish in this restaurant are negative, the rank is -1, and if all the reviews are positive, the rank is 1, with everything else in the middle. I found this to be a very convenient way to represent an attitude towards a particular dish cooked in a particular restaurant - as soon as the cumulative rating in reviews shifts to either negative or positive side, the rank of the pair "restaurant + dish" will go either below or above zero, i.e. it will literally become negative or positive, respectively.

Then I used information from the fourth element of the above data structure to calculate the mean arithmetic sentiment polarity of each combination " restaurant + dish".

Once I had a list of all possible combinations [*restaurant, dish, average rank, mean sentiment polarity*], it was easy to estimate the average rank / sentiment of one dish in the reviews of all the restaurants of the cuisine or the average rank / sentiment of all dishes for one particular restaurant. This was done by calculating the mean arithmetic value in both cases. Needless to say that I selected only the restaurants that serve Italian food by looking in the categories list of the corresponding json file.

2.3 Visualization

I chose Tableau for visualization in this task because it easily lets you change data views of the same data set by selecting and reordering different parameters. This approach turned out to be effective because I was able to receive some important insights which, otherwise, would have been hard to get based on a single data view as shown in the example in the Task 4 and 5 Instructions. When visualizing my results, I used cutoffs for the number of times a dish is mentioned to skip infrequent dishes and restaurants as they would not be interesting for analysis.

3. Results

Below are different views of the results that I obtained using my ranking system. The visualization is made in Tableau for approximately 100 top dishes or 100 worst dishes. In each figure, there are two side-by side bar charts, one for dishes vs. their rating (based on the formula mentioned above) and one for dishes vs. the number of times they are mentioned in reviews. The logarithmically scaled axes for rating and count can be seen at the bottom of the charts. The sentiment data is denoted with color in accordance with the color legend shown at the top right corner of each figure in which red means better sentiments (hot), and blue means worse sentiments (cold). None of the sentiment values are negative in this Tableau filtered subset (but I did see negative sentiment in the full set of 500+ dishes).

I also published these results in a workbook on the Tableau Public website - each figure is in a separate tab and you can use the so called tool tip which, when you move your mouse over a bar in any bar chart, shows compact numeric parameters about this particular dish. In addition, you can scroll through results for several hundred dishes instead of just 100 dishes shown in this paper and you can also sort data differently by Rating, Count, or Dish than shown in each view. Note that each data view (tab) is a Tableau filtered subset of the entire data cube of results, and these subsets may be slightly different.

The Tableau Public workbook can be found here:
https://public.tableau.com/profile/andrew.nedilko#!/vizhome/Task4and5Visualization1/ByRating?publish=yes

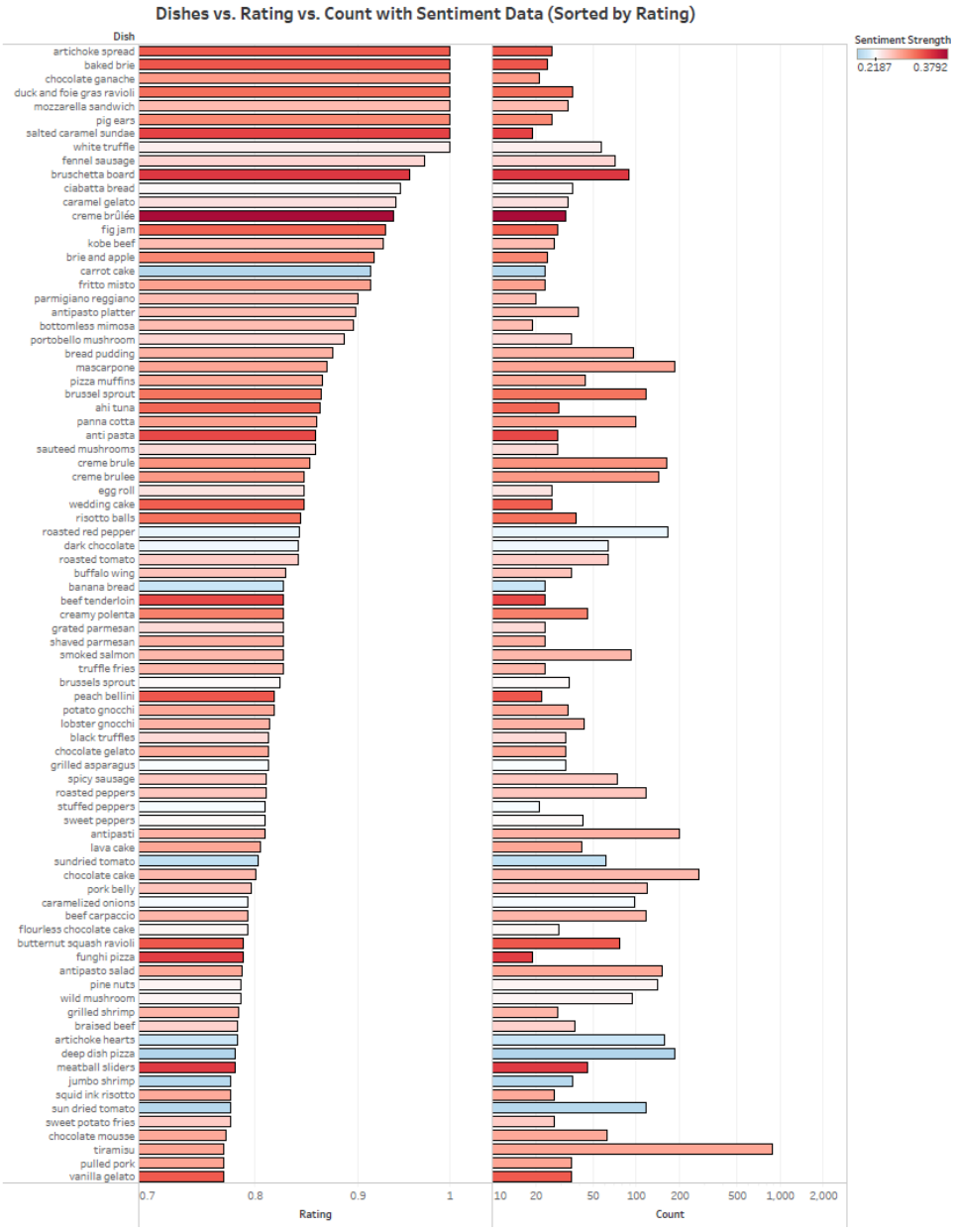

Figure 1. Dishes vs. Rating vs. Count with Sentiment Data (Sorted by Rating)

Here you can see the top dishes according to the my ranking function. The Sentiment Strength color legend is skewed because Tableau shows only a filtered subset of data here, but it is normal on the chart without filtering (for all 500+ dishes). Most of the dishes are Italian, although I never heard of pig ears. There are also some general dishes; they just happen to be served in Italian restaurants, and apparently they are good. What you can clearly see from this chart is that when the rating of dishes decreases, so does the sentiment polarity, and you can observe less intense shades of red, white, and even blue color at the bottom. There is definitely a correlation between the rating (rank) of a dish and the sentiment of reviews. To me, this is a confirmation that I chose the correct ranking function.

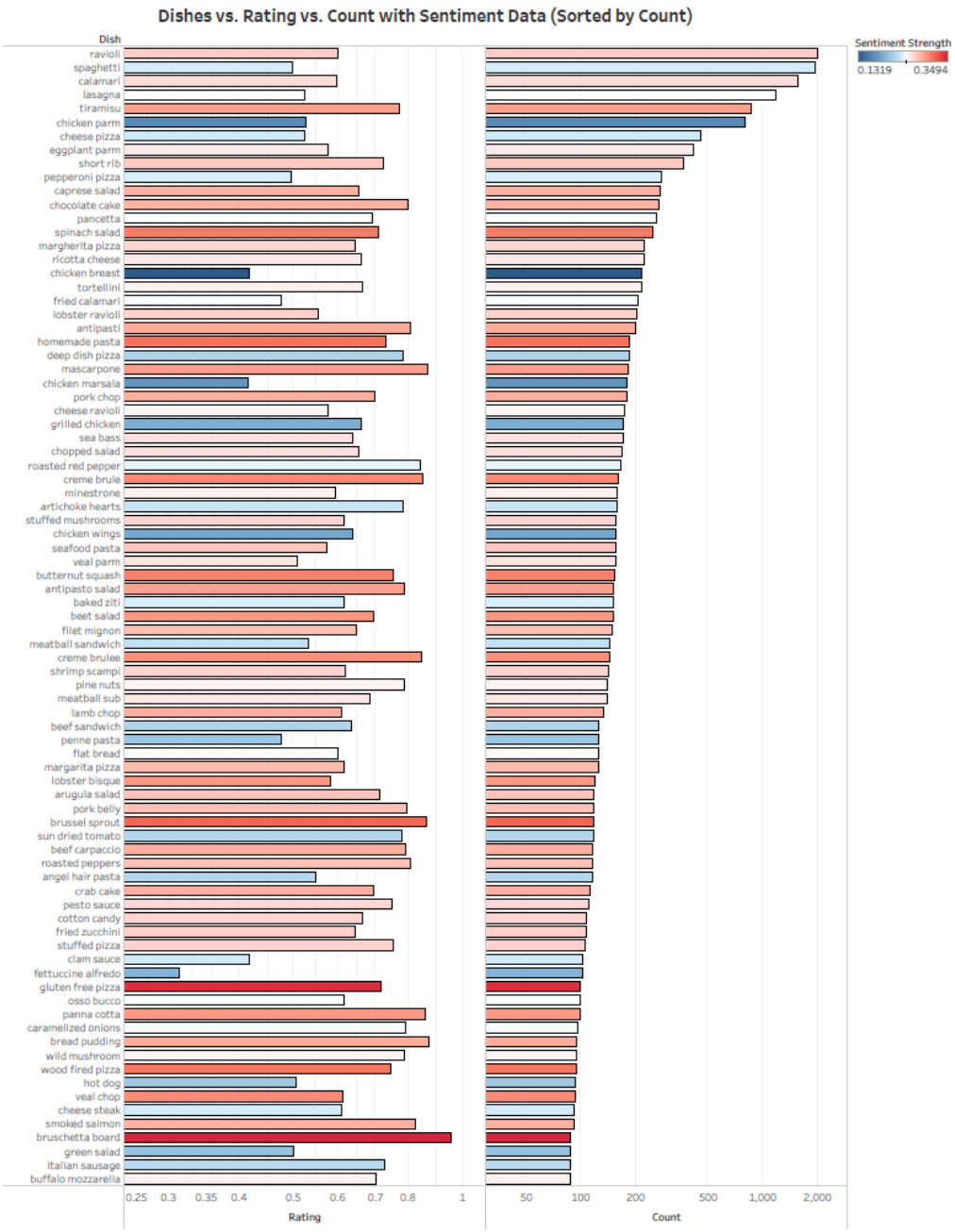

Figure 2. Dishes vs. Rating vs. Count with Sentiment Data (Sorted by Count)

Here you can see the top dishes sorted by the number of times they are mentioned in all reviews for the Italian cuisine. There is definitely little to none correlation between the count of a dish and the sentiment of reviews which means that, although such top dishes as ravioli, spaghetti, calamari or lasagna are mentioned in thousands of reviews, the customers are not always happy about them. Thus, to answer Question 1 from the first page of this report: recommending dishes based on the right ranking function possibly combined with correct sentiment analysis, is definitely a more efficient method than simply counting the number of reviews the dish is mentioned in. Logically, this should be also true about the number of times a dish is mention in the reviews for one restaurant - it is not the quantity, but the quality that determines the popularity of a dish. Therefore, one should always choose to analyze the rank and sentiment rather than the simple count, whether for the entire cuisine or for only one restaurant.

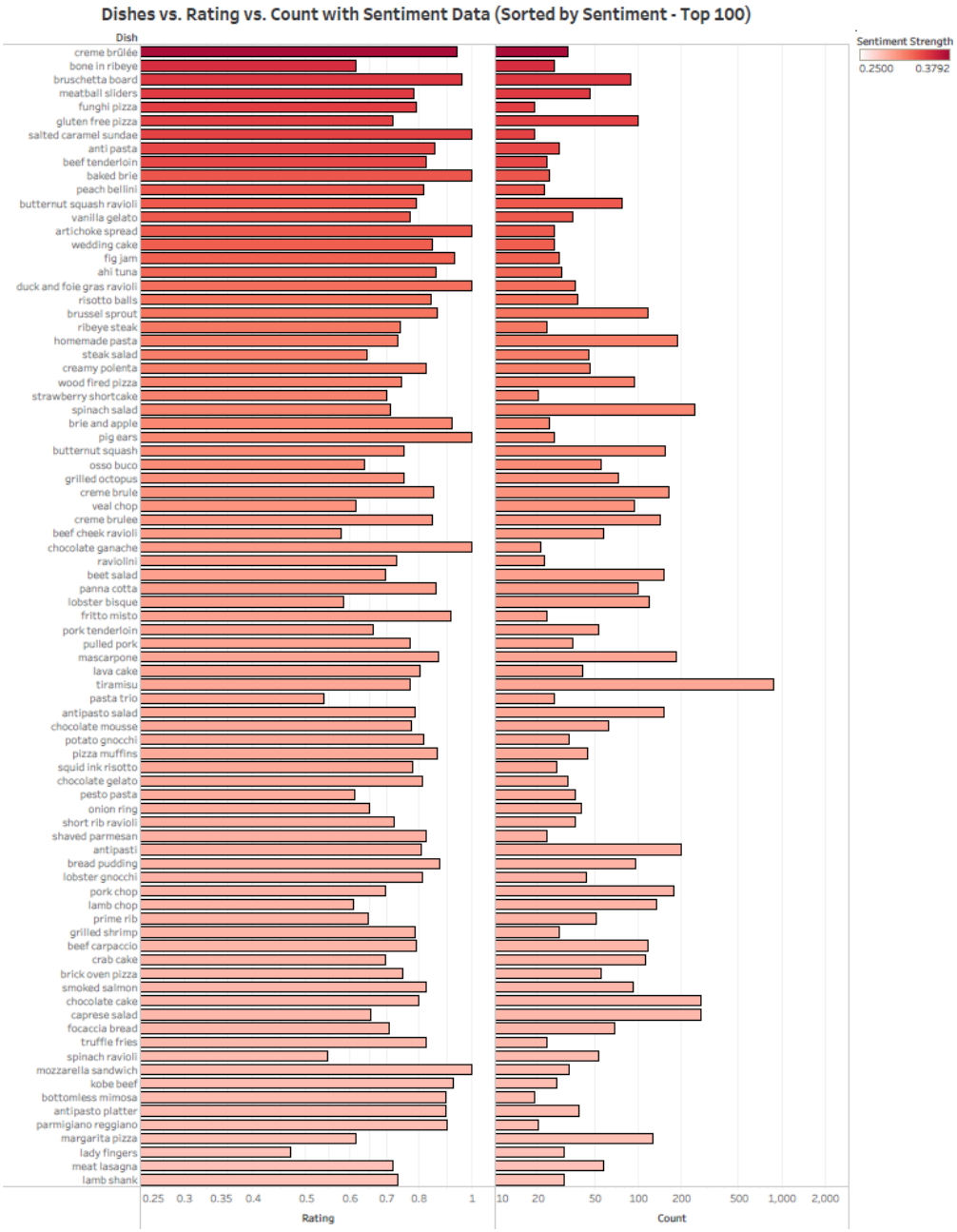

Figure 3. Dishes vs. Rating vs. Count with Sentiment Data (Sorted by Sentiment)

This chart demonstrates the top 100 dishes based on the sentiment analysis. I provided it for information purposes. More interesting results can be found on the next chart which is a reverse analogue of this one.

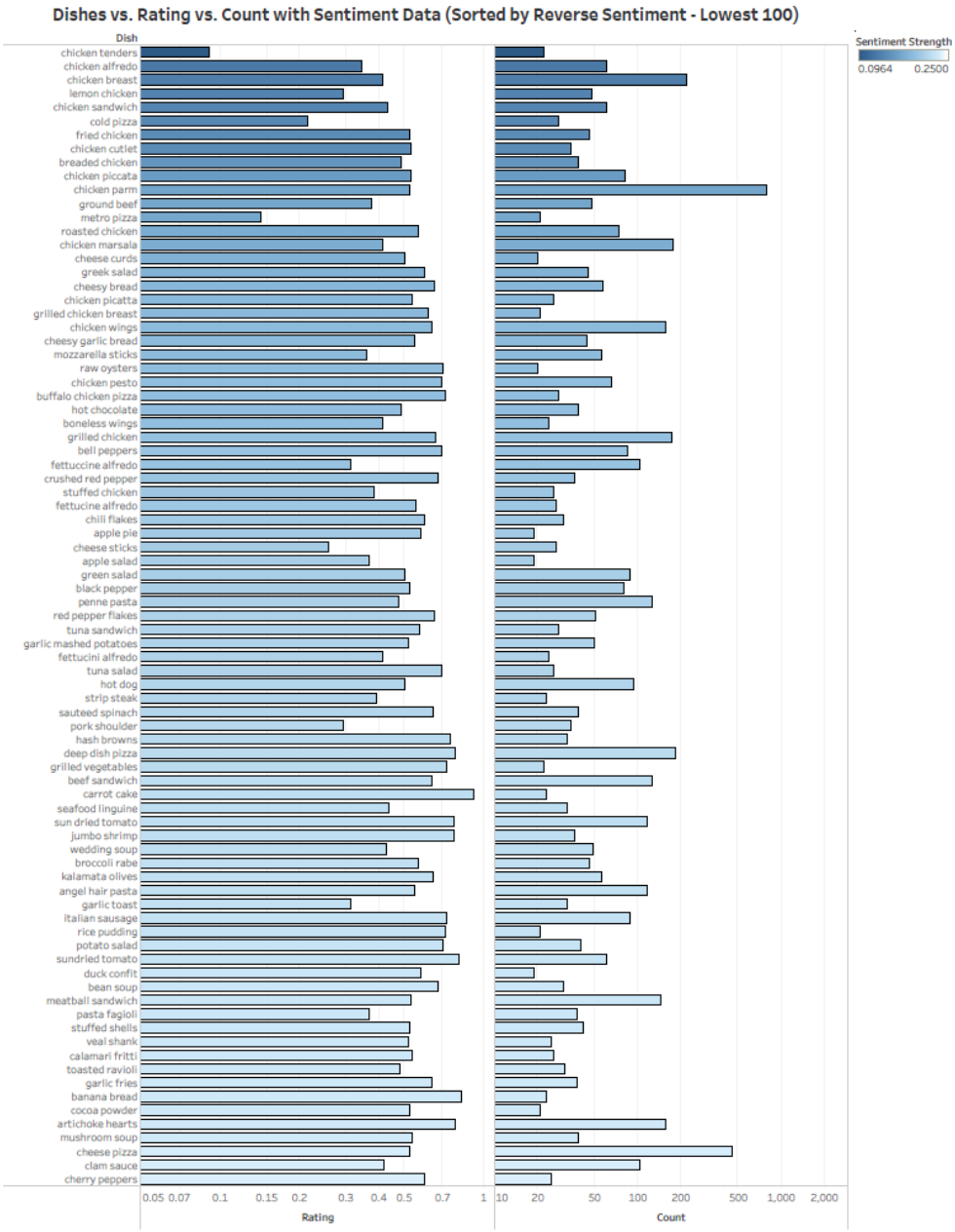

Figure 4. Dishes vs. Rating vs. Count with Sentiment Data (Sorted by Sentiment)

These are the worst dishes of the Italian cuisine with the worst ones shown in dark blue at the top of the chart. The interesting observation is about the type of the worst dish. Out of the first 10 top worst dishes, 9 have chicken in them. There is also a lot chicken in subsequent dishes. Therefore, when going to an Italian restaurant, one should probably avoid ordering chicken. Italian chefs either do not know how to or do not like to cook this kind of meat.

Below are the results for top 100 restaurants by their rank. I haven't provided several data views here because they are not so interesting as in the case with dishes. The data display and color codes are the the same. There is also an evident correlation between the rating and sentiment. A published Tableau Public workbook with more restaurants in it and more interactivity can be found here: https://public.tableau.com/profile/andrew.nedilko#!/vizhome/Task4and5Visualization2/ByRating?publish=yes

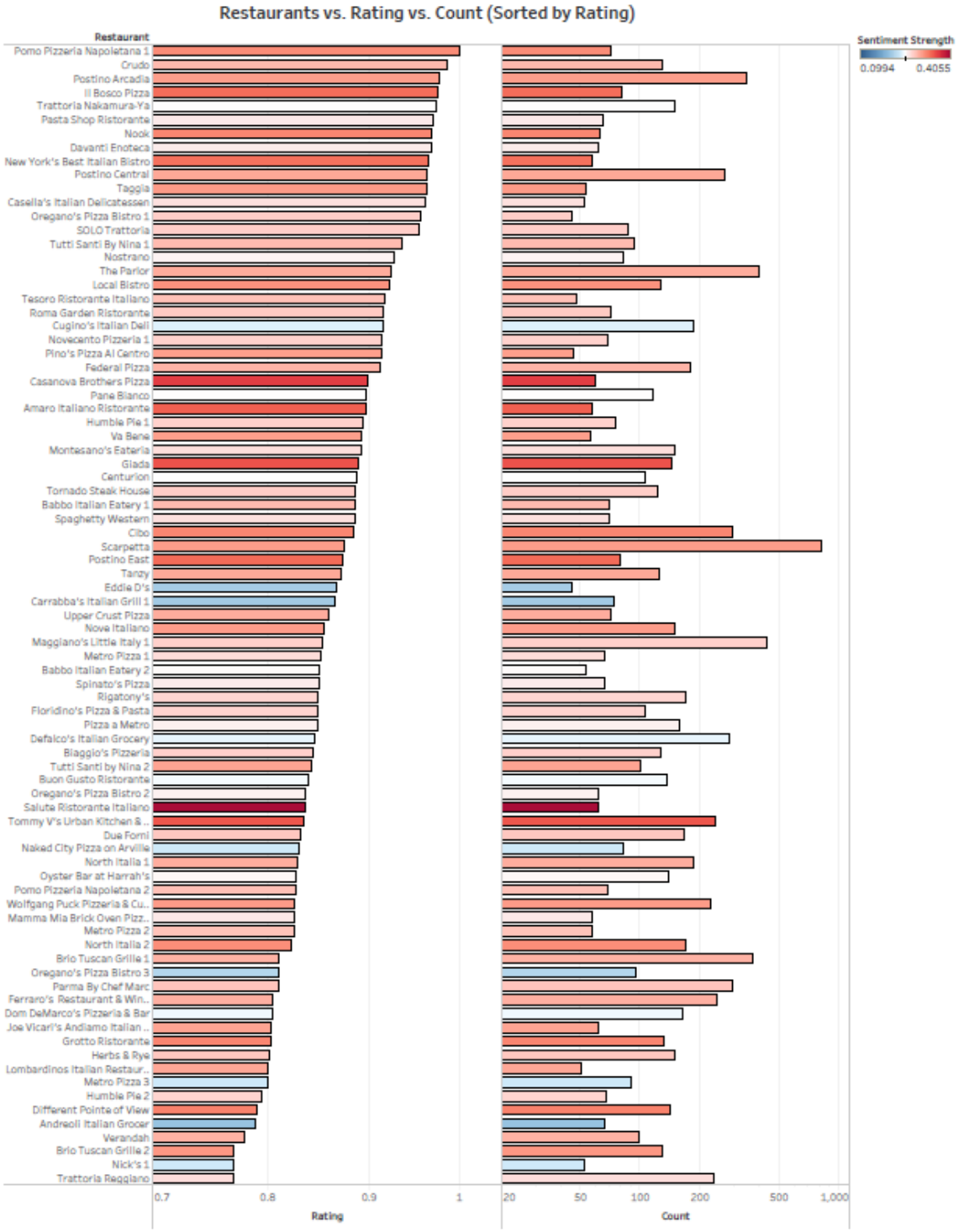

Figure 5. Restaurants vs. Rating vs. Count with Sentiment Data (Sorted by Rating)

An improvement of my results for top dishes could be accounting for how many restaurants cook a certain dish rather than using a count for the entire cuisine, and an improvement of my results for top restaurants could be extracting the location information along with the restaurant names. I didn’t do this in the interest of time, but it can be easily done using my ranking and information extraction algorithms. The location information may be a critical parameter in a project when, for example, you want to rank restaurants in a specific city or state.

4. Conclusion

Overall, I can say that my ranking function yielded meaningful results because it definitely correlates with the sentiment analysis. This was established in the process of analyzing several data views of the same data set which turned out to be more efficient than having just one chart.

The results obtained also proved that, when ranking and recommending dishes and restaurants, quality is more important than quantity. So, first one has to analyze ranks/ratings and the sentiment polarity, and only then use count as a secondary parameter, if needed.