CAPSTONE PROJECT FINAL REPORT: RECIPE SENTIMENT ANALYSIS AND RECOMMENDER SYSTEM

Antonio Gagliardo | June 25, 2023

Problem Statement and Background

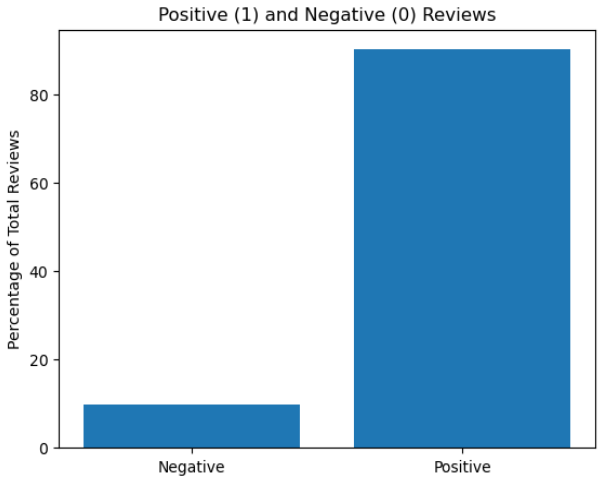
The problem that is addressed in these projects is the challenges that individuals have ingredients and they have no idea what to do with them. People like me who are not very good at calculating amounts when making food often have some leftover ingredients that can get rotten if we don't use them fast. Also, companies that have a business model that revolves around food or recipes sometimes struggle with what to recommend to their visitors/customers to retain them. The first part of this project focuses on a sentiment analysis to understand the customer better, know their sentiment and what they value and what they dislike. This is useful cause making it with a model is time effective and can be reused multiple times, while doing manually can prove time-consuming and costly. The next part of the project will aim to develop two different recommendation systems, one based on content (recipe/ingredients) while the other one will be user-based (reviews).

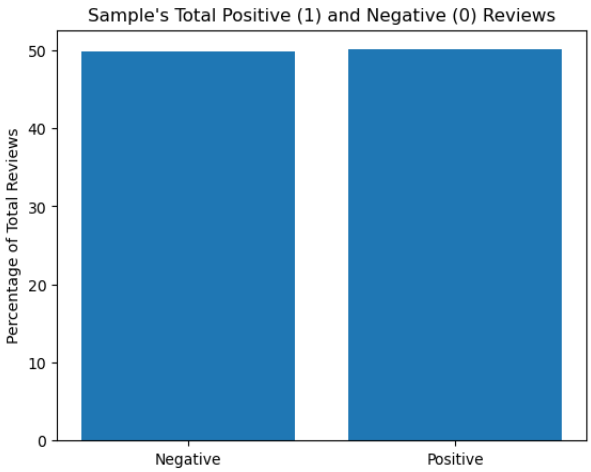
About the data

The dataset was obtained from Kaggle and consists of around 500,000 recipes and 1.5 million reviews from the website Food.com (formerly known as Zaar). There are 2 datasets, the first dataset includes information such as recipe details (author, cook time, nutritional value, aggregated ratings, ingredients, description, etc.), and the second dataset is the review details (author id, date, review itself, review score, etc.) and other relevant attributes.

Data Cleaning and Exploratory Analysis

Before proceeding with the analysis, the data was cleaned. Some steps included getting rid of null values, taking care of data imbalance in reviews (as shown below, the before/after up-sampling)





Balancing was important because it can create a bias in our model when it wants to predict if a comment is positive or negative.

Exploratory analysis was done to gain insights into the data set. Such as reviewing the distribution of recipes across categories, analyzing ingredients usage, and exploring review sentiments.

Modeling

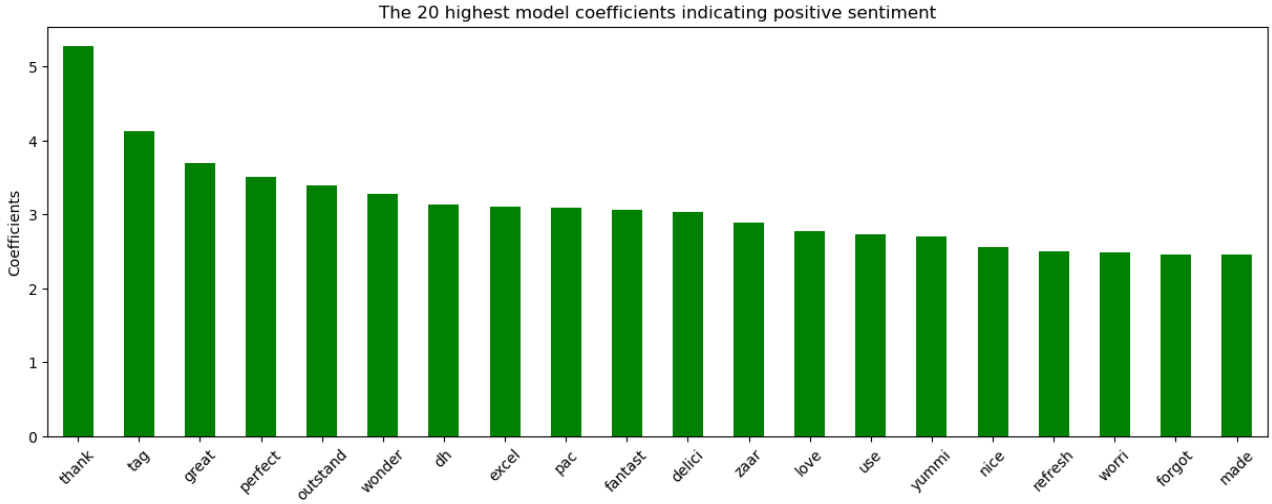
For the sentiment analysis, a text mining/tokenizing approach was adopted using the Count Vectorizer and TF-IDF vectorizer (which breaks sentences in a numeric way to help my model learn) and with the help of a logistic regression, I could get very valuable data and insights. Additionally, KNN and decision tree models were applied to evaluate the accuracy of different approaches.

In the content-based recommendation system, I also used the TF-IDF vectorization technique but applied it to the ingredients part. By using this vectorizer, the ingredients were transformed into numerical features (vectors), and in conjunction with the 'cosine similarity', which allows computing how similar the items (in this case recipes) are, I could retrieve the recipes that are the most similar to the recipe on the input.

Finally, on the user-based recommendation I had to put the users and recipes in a matrix and do a function based on 'cosine similarity' to compare all the users in the matrix and see users similar to the one I want to predict and make an educated prediction on if this specific user will like or not the recipe based on the review of similar users.

Results

In the sentiment analysis phase, I conducted an in-depth analysis to see which words are strongly associated with positive reviews, and which ones are associated with negative reviews.



This helped as a base to use coefficients to accurately predict sentences. At the end of that phase, I made a function that was able to discern the positive and negative sentiments accurately. I tried with a few sentences and it was correctly predicting the sentiment of the text that I was inputting.

Moving to the content-based recommendation system, I got to train it in a similar way to the sentiment analysis, but training the model on the ingredients and keywords. Also did a function for this one, and the functionality enables the user to explore and discover new recipes that align with their preferred ingredients. An example when searching “Warm Chicken A La King”:

	RecipeName	similarity	Ingredients
2	Warm Chicken A La King	1.000000	c("chicken", "butter", "flour", "milk", "celer...
16149	Classic Chicken Ala King	0.624533	c("butter", "flour", "salt", "pepper", "chicke...
16772	Garlic Mushrooms With Basil	0.536586	c("butter", "garlic cloves", "button mushrooms...
21231	Quickie Tom Yum Soup	0.516537	c("garlic", "celery", "tomatoes", "button mush...
6694	Cajun Glazed Mushrooms	0.508901	c("button mushrooms", "unsalted butter", "marg...

Finally, the user-based recommendation system showed potential through the predict rating function that I made at the end. This function could accurately estimate the rating a current user would assign to a recipe. With this last integration, this project paves the way for the development of a practical and tailored food recommendation system.

Conclusion

The food recommendation system that was done in this project shows potential since it solves the problem of individuals having ingredients but lacking recipe ideas. By implementing sentiment analysis, valuable insights can be extracted from customer reviews, making it valuable for businesses to better understand customer preferences.

Overall, the project demonstrates promise in offering insights and recommendations that are useful for users and businesses.