Cooking the Model: Predicting Ratings from Recipe Reviews with NLP

Garrett Allen and **Nathan Ostrowski** and **Alex Bildner**
Duke University

Abstract

Online recipe platforms offer an extensive repository of user-generated reviews and ratings that help consumers decide which recipe to cook and how to modify it to suit their preferences. We present a replication study of the 2019 paper "My Curiosity was Satisfied, but not in a Good Way": Predicting User Ratings for Online Recipes by Can Liu et al., which predicted recipe ratings using NLP techniques. Our study extends the original work by exploring various models, including Lasso Regression, Support Vector Machine (SVM), Naive Bayes, and Random Forest, with updated engines for sentiment classification and part of speech tagging, and using a new dataset from Food.com. Surprisingly, most models perform no better than the baseline, which simply predicts a 5-star rating. However, the Random Forest model shows potential with slightly lower accuracy but a lower root mean squared error. Our study contributes to the understanding of the strengths and limitations of various NLP techniques for predicting recipe ratings from user-generated reviews and offers directions for future research in this area.

1 Introduction

Online recipe platforms present one of the internet's largest repositories of user-generated reviews and ratings. When choosing a recipe, these ratings and reviews serve an incredibly important role for consumers—empowering them to confidently choose which recipe to cook and how to modify it to fit their tastes. Recipe reviews are also a goldmine of natural language information—varied, complex, often highly personal, and often balanced between praise/criticism of a base recipe and suggestions for tweaks and improvements. In this paper, we present a replication study of the 2019 paper "My Curiosity was Satisfied, but not in a Good Way": Predicting User Ratings for Online Recipes by Can Liu et al., which sought to predict recipe ratings from user reviews using natural language processing techniques. Our study extends the original work by exploring a variety of models (including Lasso Regression, Support Vector Machine (SVM), Naive Bayes, and Random Forest) along with updated engines for sentiment classification and part of speech tagging and a new, rich dataset.

While the original study uses computationally expensive n-grams with 6,000 features to power their predictions and only briefly wades into the predictive power of lexical/linguistically motivated features, we seek to complement their work by further exploring the value of lexical/linguistically motivated features for our analysis. In contrast to the original study which focused on SVMs, we explore the effectiveness of many types of predictive models. In addition to adding model types, we update the engines used for Part of Speech tagging (English EWT), Sentiment analysis (NRC), and feature extraction. We also iterate upon the original study by using a different dataset (Food.com recipes and their reviews) to extract relevant features.

Our findings indicate that most models—even with updated engines for part of speech tagging and sentiment analysis—perform no better than the baseline, which simply predicts a 5-star rating. The Random Forest model however shows some promise, with slightly lower accuracy than the baseline but a lower root mean squared error. This study contributes to the understanding of the strengths and limitations of various NLP techniques for predicting recipe ratings from user-generated reviews, and offers directions for future research in this area. Our findings also indicate that while the original dataset is heavily skewed toward 5-star ratings, upscaling the training set does not increase the accuracy of the model. In-fact, it decreases it (from 72% to just 31%).

2 Literature Review

Food and diet are immensely important facets of cultures all around the world and perhaps the single most important indicator of health and well-being. Thus, it's no surprise that the field of NLP and more broadly machine learning research relating to food and recipes is strong. In this review, we'll cover a sample of recent influential literature in the space.

Freyne et. al (2011) explores the use of food recommender technologies, looking to help users get into good diet and exercise habits. By understanding user reasoning, Freyne et. al. (2011) look to find ways to exploit this understanding in the recommendation process, using information about how a user thinks to provide them with healthy recommendations that stick. The study implements three personalized recommender algorithms: two standard recommender strategies and one machine learning strategy. Interestingly, the authors find that users in their dataset tend to reason from food $\rightarrow$ cuisine type (affection for beef $\rightarrow$ interest in Chinese food) rather than cuisine type $\rightarrow$ food (I like Thai cuisine $\rightarrow$ affection for peanuts/noodles). Their results further show that understanding user reasoning can enhance the power of recommender algorithms and highlights the potential for future research in areas such as informative rating acquisition, item diversity, and persuasive techniques (Freyne et al., 2011).

Trattner, Moesslang, and Elsweiler's (2018) study on the predictability of online recipe popularity complements our own paper on recipe rating prediction by demonstrating the correlations between recipe features and proxies for popularity. The authors compared the Allrecipes.com and Kochbar.de platforms to investigate how to identify popularity patterns in different online food communities. They use the number of comments and ratings as indicators for popularity and find a set of features encompassing recipe content, presentation, nutrition, healthiness, complexity, seasonality, and innovation to explain popularity. Their study finds that recipe 'innovation' was the most important factor for users on Allrecipes.com, while user behavior leaned more critical on the site Kochbar.de, and attributes of uploaded recipe images were a strong signal of rating strength on both platforms. Overall, the study demonstrated the ability to predict recipe popularity to a high degree using associated metadata and provided a strong platform for future research in popularity prediction in online recipe settings (Trattner et al., 2018).

In a somewhat orthogonal exploration of machine learning and recipes, Chen and Ngo (2016) from the University of Hong Kong investigate the effectiveness of using deep learning to find recipes from a given image. Similar to other papers covered in this literature review, the authors view this primarily through the lens of health and nutrition, investigating the feasibility of giving users access to nutrition information for any food they eat. Their food dataset specifically covers Chinese dishes, and their recipe data comes exclusively from a Chinese recipe site. The authors find that deep features perform significantly better than hand-crafted features when identifying recipes, but the task of identifying recipes from mere images still remains difficult. The authors note that the way a dish is cooked, in particular, has a striking influence on lowering the accuracy of their models. The way that certain ingredients are cut (e.g. chopped, sliced, minced, cubed) also proves to be a major inhibiting factor in the development of accurate models. (Chen and Ngo, 2016).

Another influential study investigates ingredient pairings – which involves looking to see if models can identify patterns across common ingredient pairings in food and drink and are then able to predict whether unseen ingredient pairs would pair well or poorly (Park et al., 2019). The researchers used deep learning to train the model which made their predictions. Their model showed great promise, as it recommended certain pairings which are known to exist in various cuisines across the world, despite never appearing in the model's training data.

Another relevant topic to our project is the authenticity of reviews. While one of the benefits of the internet is that anyone can write a review and it is essentially free to read and write reviews, one of the downsides of the internet is also that, well, anyone can write reviews for essentially free! This leads to issues with fake reviews. One particularly interesting paper on the topic of how to detect fake reviews is "Fake review detection on online E-commerce platforms: a systematic literature review" (Paul and Nikolaev, 2021). While their paper described the multitude of different methods and techniques wielded to attempt to identify phony reviews, they concluded that the reality of the situation was that no method or model has been able to truly solve this problem completely, as the

task itself presents numerous challenges, including but not limited to the lack of scalability of certain model detection architectures, the domain specificity of many successful approaches (not generalizable), and how highly equipped and motivated those propagating fake reviews are, making it difficult to differentiate their fake reviews from real ones (Paul and Nikolaev, 2021).

A frequently studied topic in Natural Language Processing is sentiment analysis. Relevant to our topic, many interesting studies attempt to do sentiment analysis of reviews. One interesting study on this topic is "Using Machine Learning to Predict the Sentiment of Online Reviews: A New Framework for Comparative Analysis" (Budhi et al., 2021). This paper features a series of experiments and different methods to predict the sentiments of reviews, using review-star rating pairs as training data. The study concluded that it is possible to determine positive or negative sentiments of reviews with a reasonable level of accuracy, getting more granularity, specifically adding predictions for neutral reviews to the output, is more difficult (citation). They also demonstrated that there is a limit to how much benefit is to be gained from adding more training data and more features to the model, specifically noting "[o]ur experiments indicated that 5000 features and 500,000 reviews are the cut-off points for polarity prediction" (Budhi et al., 2021).

3 Methods

In our research, we follow the results of two papers on predicting recipe reviews. We pursue a replication study of the results of (Liu et al., 2014), where they attempted to predict the ratings of a review given information about a recipe (reviews, ingredients, etc). Following the work of (Yu et al., 2013), we use the reviews of our recipes to predict their star rating on a 1-5 star rating basis, as they showed that ingredients did not help when predicting the quality of recipes. We create a two-stage classifier, where we first try to predict individual review ratings, and then from that, try to aggregate those reviews to produce an overall rating for a dish. While the authors of (Liu et al., 2014) were not able to find lexically and linguistically motivated features that would performatively predict ratings from review data, our study seeks to build upon this work by focusing squarely on these features, seeking to update the features, classification engines, dataset, and models used. Unlike previous works in (Yu et al., 2013), (Liu et al., 2014), we try out multiple different classifiers besides SVMs to compare their performance (e.g., random forest, naive bayes, lasso, and a baseline model). We also create different classes than in (Liu et al., 2014), which they cite as a further direction for their work. We apply our work to a different dataset from the website Food.com, rather than the Epicurious dataset that they had been working with. [1]

3.1 Feature Engineering

In the process of preparing our data for modeling, we engineered a range of linguistically and lexically motivated features. Building on the work of (Yu et al., 2013), some of these features include:

1. The mean percentage of personal pronouns per sentence (which required first doing part of speech tagging using the English EWT engine).
2. The mean number of words per sentence.
3. The total number of words in the review.
4. The percentage of passive sentences per review.
5. The mean number of punctuation marks per sentence.
6. The mean number of capitalized characters per sentence.
7. The type/token ratio per review.
8. Sentiment analysis of reviews using the NRC sentiment engine.

Explanations of why these variables were chosen can be found in (Yu et al., 2013), but broadly, they were chosen because they were thought to model the intensity of user reactions and the positivity/negativity of their sentiments.

3.2 Models and Results

We compared the performance of five different models: a baseline model, Lasso, Support Vector Machine (SVM), Naive Bayes, and Random Forest. The baseline model was designed to predict

[1] Our dataset can be found here: https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions?select=RAW_interactions.csv

a 5-star rating for all instances (as the majority of the recipes received 5-star ratings) and served as a point of comparison for the other models. For the SVM model, we experimented with both radial and linear kernels, finding that the kernel did not matter—the model still aligned with the baseline. In the case of the Random Forest model, we explored the impact of using different numbers of trees (1,000 and 3,000), finding that the number of trees beyond 1,000 did not make a significant difference in accuracy. The following figure gives an overview of model performances using our original training set:

Table 1: RMSE and accuracy on out of sample test set

Model	RMSE	Accuracy
Baseline Model	1.493	.7222
Naive Bayes	1.622	.6726
Lasso Regression	1.493	.7222
SVM	1.493	.7222
Random Forest	1.491	.7217

We observe in Table 1 that the Lasso Regression, SVM, and Baseline models all performed with exactly the same accuracy. There's a simple reason for this: all three predicted 5-stars on every recipe, matching the fact that 72% of the dataset's ratings were 5-star (see Fig. 1). We observe in the predictions of our Random Forest model that not every rating given is 5-star, thus its near-parity performance with SVM, Lasso, and our Baseline model is much more impressive. It also very slightly improves on RMSE, suggesting that even if it was less accurate it was very slightly more likely to get closer to the real rating.

Because our range of models tended to align with the baseline strategy of naively predicting 5-stars on every recipe, we decided to upscale our training set to see if we could induce interesting behavior in our models. Upscaling is a statistical technique where the minority classes (in this case, 0-4 stars) are duplicated in the training set to have as many observations as the majority class (5 stars). The idea is that, by inflating the minority classes, we might amplify the trends found in the minority classes so that the models better detect these observations. The method did not improve performance; in fact, after upscaling the training set, all of our models performed significantly worse at the task of predicting ratings in the test data. In fact, the

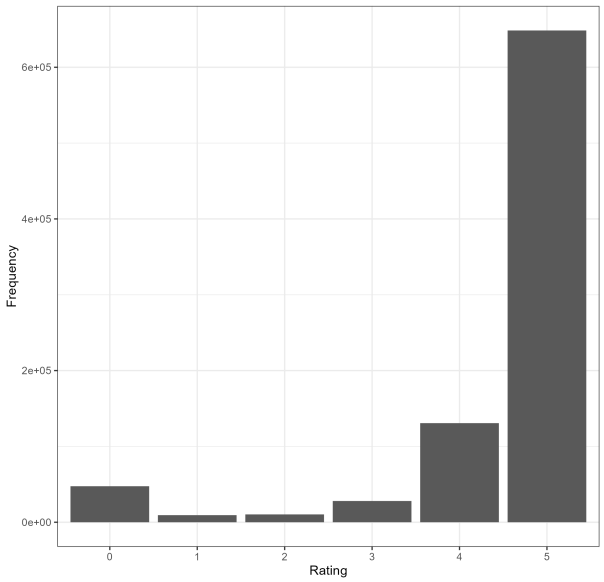

Figure 1: Breakdown of original training set reviews by rating

accuracy of the SVM dropped from an initial 71% (predicting 5-star for every review) to 31%. The only model from our set to perform well on this updated dataset was Random Forest, which yielded an accuracy of 69.9% using 1,000 trees (note that this is lower than training the original training dataset, where the Random Forest model yielded 72% accuracy). We tested this Random Forest model further on 30,000 trees, finding that the accuracy increased only very slightly to 70.1%, still below the accuracy of the Random Forest model trained on our original dataset.

Further investigating how to modify our Random Forest model to increase its performance, we then attempted to construct six different Random Forest models, each focused on predicting a particular star rating (i.e. specific 0-star, 1-star, etc. models). We noticed that this 6-model classification approach actually performed worse than the base Random Forest model (70% accuracy).

4 Discussion

Our study aimed to replicate and extend the work of Can Liu et al. (2019) in predicting user ratings for online recipes using various natural language processing techniques. Our work diverged from the original study in several key aspects, such as exploring a number of different predictive models, using updated engines for part of speech tagging and sentiment analysis, and analyzing a different dataset from Food.com.

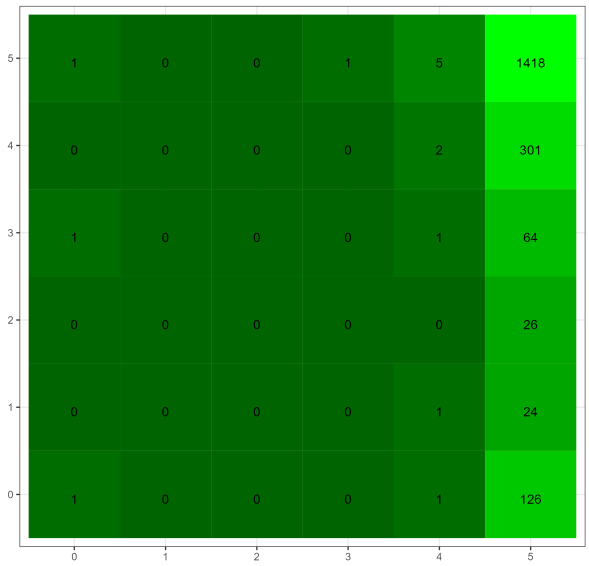

Figure 2: Confusion Matrix of Random Forest Model

A surprising outcome of our analysis was that most of the models, even with updated engines for part of speech tagging and sentiment analysis, performed no better than the baseline model, which simply predicted a 5-star rating for all instances. This could be attributed to the heavy skew towards 5-star ratings in the dataset, which might have led the models to adopt a conservative approach in making predictions. It is important to note, however, that the Random Forest model showed some promise—exhibiting slightly lower accuracy than the baseline model but with a lower root mean squared error. This indicates that the Random Forest model was more likely to provide predictions closer to the actual ratings, despite being less accurate overall.

We also found that upscaling the training set did not lead to better accuracy. In fact, the accuracy of most models, including the SVM model, decreased significantly (from 72% to 31%). This finding contradicts the conventional wisdom that larger training sets generally improve model performance, and the conventional wisdom that a balanced dataset ought to lead to better model training. A potential reason for this decline in performance could be the introduction of noise or other confounding factors in the enlarged dataset, which may have made it more challenging for the models to discern patterns in the data.

Another interesting finding from our research was the differing impact of various linguistic and lexical features on model performance. Some features, such as the number of words per sentence, impacted rating prediction significantly (though it's difficult to know the exact impact, as the models used still predicted 5-star ratings nearly every time). This highlights the importance of feature selection and engineering in building accurate and efficient models, and certainly represents an important focus area for future work.

Our research contributes to the understanding of the strengths and limitations of various NLP techniques for predicting recipe ratings from user-generated reviews. While our findings may not provide a comprehensive solution to the problem, they offer valuable insights and directions for future research in this area.

Specifically, future research in this area could explore other feature sets, including features derived from the recipe content itself—such as ingredient pairings, ingredient quantities, cooking techniques, or recipe complexity. It would also be worthwhile to investigate other models (deep learning techniques in particular) for predicting user ratings. Additionally, addressing class imbalance in the dataset, either through techniques such as oversampling or undersampling, or by exploring alternative evaluation metrics that are less sensitive to class imbalance, may lead to improved model performance. Finally, given that our work assumed each review given was written by a human, future work could explore the impact of fake reviews on the predictability of recipe ratings and devise strategies to mitigate their influence.

5 Conclusion

Our study replicated and expanded on the techniques of (Yu et al., 2013) to investigate the usefulness of lexically and linguistically motivated features using numerous NLP approaches to predict ratings from recipe reviews. While some of our models (SVM, Naive Bayes) performed no better than the baseline, the Random Forest model demonstrated potential for further research. Furthermore, while upscaling significantly decreased the accuracy of most models, the Random Forest model was only marginally negatively affected. Future studies should explore alternative models, feature engineering techniques, deep learning techniques, and normalization strategies to improve the prediction of recipe ratings and enhance the real value of user-generated content in online recipe platforms.

References

Gregorius Budhi, Raymond Chiong, Ilung Pranata, and Zhongyi Hu. 2021. Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis. *Archives of Computational Methods in Engineering*, 28.

Jingjing Chen and Chong-wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 32–41, New York, NY, USA. Association for Computing Machinery.

Jill Freyne, Shlomo Berkovsky, and Gregory Smith. 2011. Recipe recommendation: Accuracy and reasoning. In *User Modeling, Adaption and Personalization*, pages 99–110, Berlin, Heidelberg. Springer Berlin Heidelberg.

Can Liu, Chun Guo, Daniel Dakota, Sridhar Rajagopalan, Wen Li, Sandra Kübler, and Ning Yu. 2014. "my curiosity was satisfied, but not in a good way": Predicting user ratings for online recipes. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 12–21, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Donghyeon Park, Keonwoo Kim, Yonggyu Park, Jungwoon Shin, and Jaewoo Kang. 2019. KitcheNette: Predicting and ranking food ingredient pairings using siamese neural network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Himangshu Paul and Alexander Nikolaev. 2021. Fake review detection on online e-commerce platforms: a systematic literature review. *Data Mining and Knowledge Discovery*, 35.

Christoph Trattner, Daniel Moesslang, and David Elsweiler. 2018. On the predictability of the popularity of online recipes. *EPJ Data Science*, 7(1):20.

Ning Yu, Desislava Zhekova, Can Liu, and Sandra Kübler. 2013. Do good recipes need butter? predicting user ratings of online recipes.