Value Proposition for Home Cooks vs Bakers

Created by Michael Renehan

Table of Contents

- Introduction
- Data Collection
- Hypothesis
- Sentiment Analysis
- Vocabulary Analysis
- Candidate Model Comparison
- Conclusion and Recommendation

Introduction:

A food publisher sees an opportunity for improved content and marketing strategy by connecting with the unique needs of home cooks and home bakers

Data Collection:

- Pushshift API
- /r/Cooking
- /r/Baking
- 5000 posts
- Title & Text
- Length & Word Count

Hypothesis

While cooking and baking are related activities, and all baking is technically cooking, these groups represent very different personas and user needs. Users of the cooking subreddit may need 30 minute meals for a family, guidance for particular dietary restrictions, or simply the “best” something without a lot of research.

Home bakers on the other hand tend to be hobbyists looking for their next weekend project, curious to pick up new techniques, or may simply take joy in sharing the fruits of their labor with friends and family.

By identifying unique vocabularies and classifying posts on food content, we may better meet user needs and engage with them.

Sentiment Analysis

To begin to understand if these users really relate to these topics differently, I began with sentiment analysis across about 5,000 reddit posts in the Cooking and Baking subreddits.

0.11

Positive sentiment scores in aggregate for cooking and baking scores were both about the same, 0.11. Nothing to see here.

Same positive sentiment, different needs

Selections from the top ~50 most positively-scored posts in each group

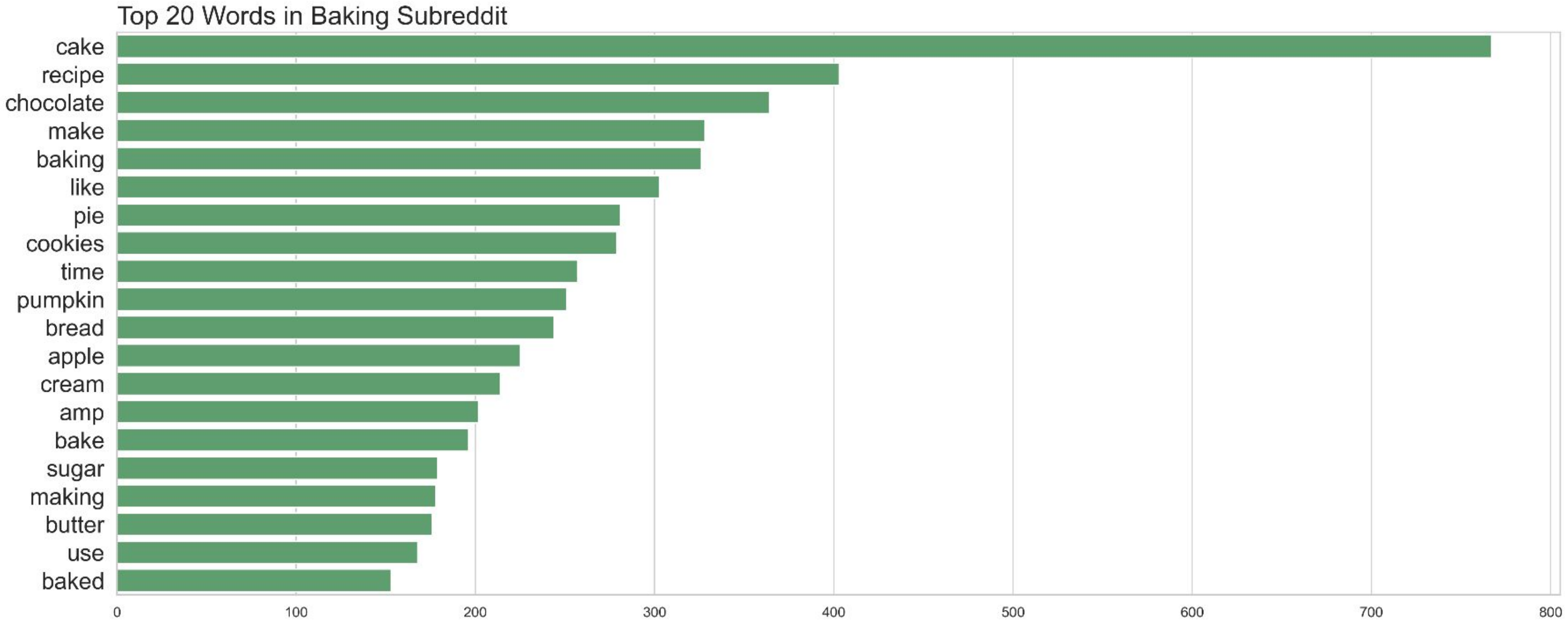
Happy Cooks Say:

- “MOST DELICIOUS RISOTTO RECIPE EVER”
- “Best High Protein Frozen Meals”
- “Best Cookie Baking Design Ideas”
- “Best Christmas Pot Roast Recipe”
- “Best stuffing for a roast chicken?”

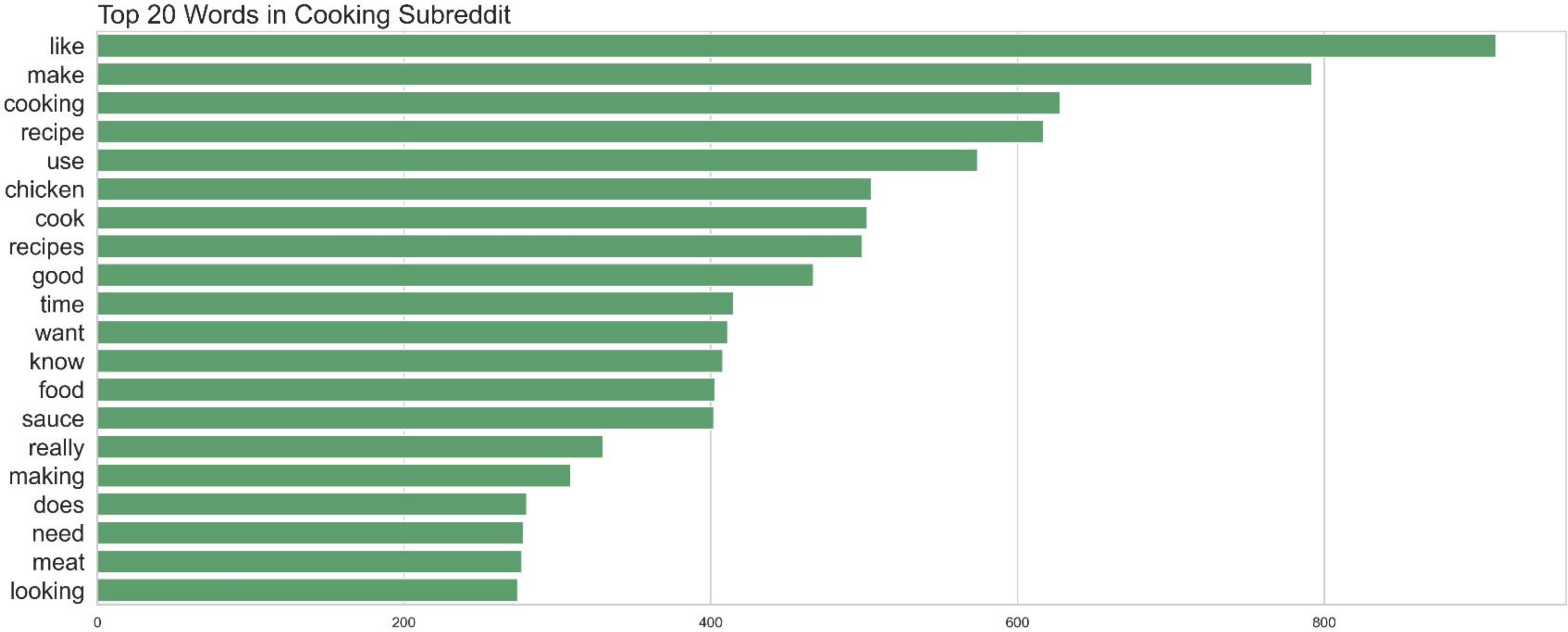
Happy Bakers Say:

- “My LOVELY cinnamon rolls”
- “Cornbread goes great with chili”
- “Super easy chocolate cake recipe”
- “Happy (Canadian) Thanksgiving!”
- “First time making sweet potato pie, it was delicious”

Vocabulary Analysis: Baking



Vocabulary Analysis: Cooking



Vocabulary Analysis

The only *common* top words

- Recipe
- Make
- Use
- Time
- Like

These distinct vocabularies support the hypothesis that these subreddits represent distinct personas and needs, and that a classification model could be successful at distinguishing between them.

Model Evaluation

Three high-level approaches were evaluated to build a classifier that effectively distinguished between posts from the Cooking and Baking subreddits, respectively, and tens of thousands of configurations were searched over to find the optimal approach.

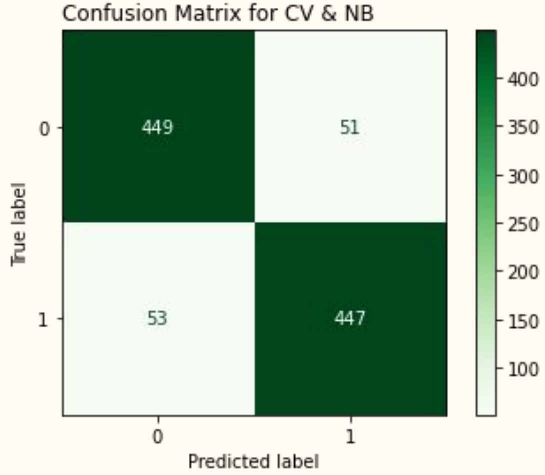
These models were selected because of their efficacy at natural language processing and the relative ease of testing many, many configurations.

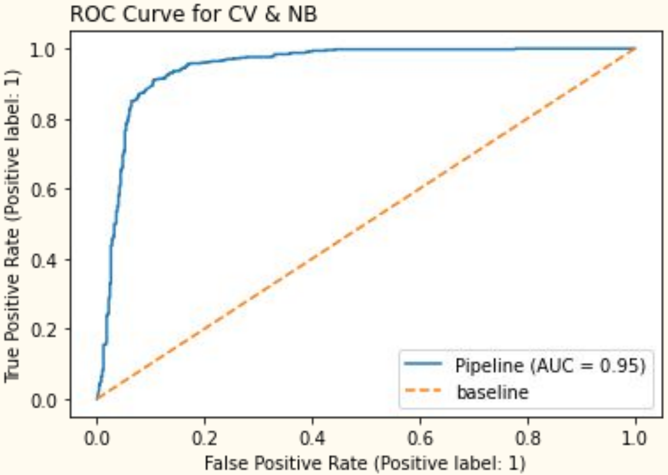
1. Naive Bayes using CountVectorizer
2. Naive Bayes using TFIDF
3. Random Forest using Count Vectorizer
4. Random Forest using TFIDF

Naive Bayes using CountVectorizer

This model, using CountVectorizer and Naive Bayes, produced a solid accuracy score of 0.895 and had the **most balanced false predictions** between the two classes.

Since baking, taxonomically, is a subset of cooking, I'm paying special attention to the balance of false predictions, since generalizing to cooking is less precise.

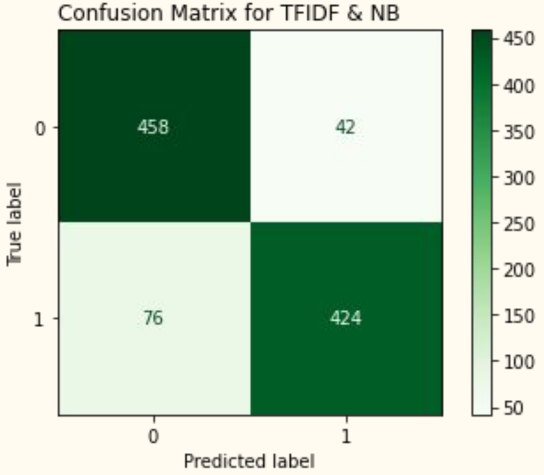


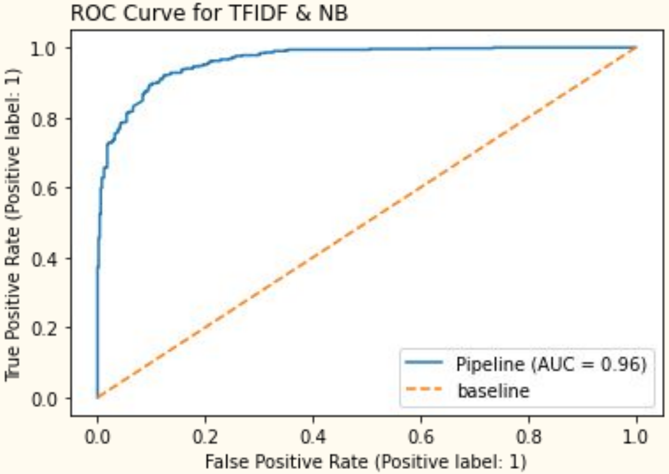


Naive Bayes using TFIDF

Next, I looked at the same model using a different approach to handling text. My hypothesis was that the distinctive vocabularies would really shine here using Term Frequency, Inverse Document Frequency.

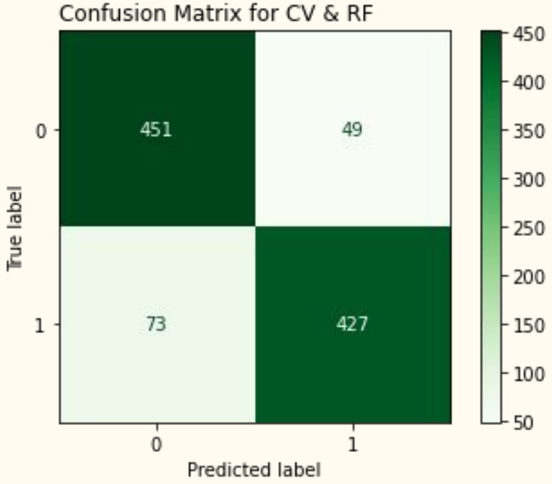
However, the test accuracy score (0.882) was about the same as before, with many more false negatives in the Baking category, so this approach was deemed inferior to the previous one.

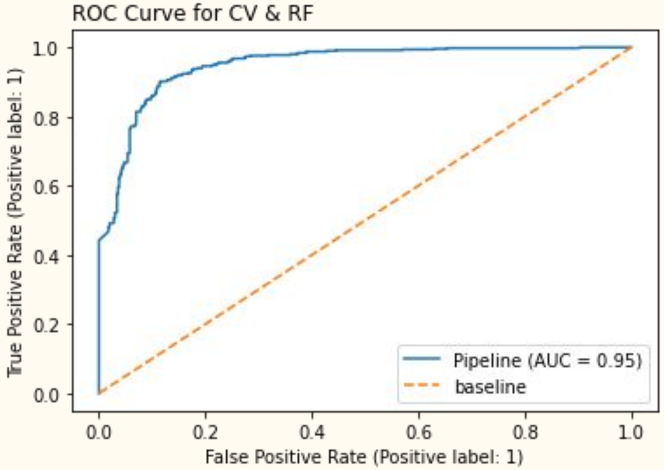




Random Forest using CV

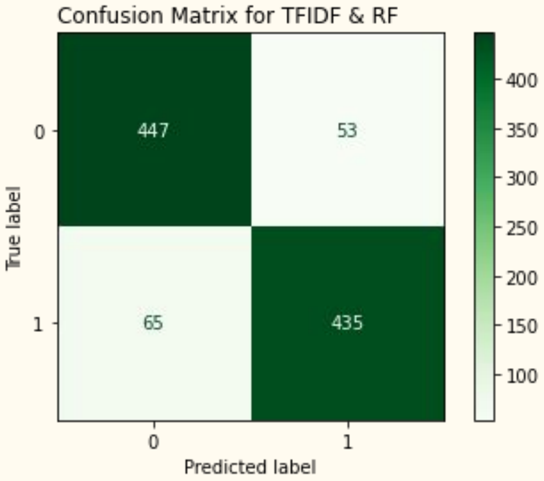
The best gridsearch parameters for Random Forest using Count Vectorizer performed very similarly to the previous configuration: similar accuracy score, but again many more false positives for baking, which it our target.

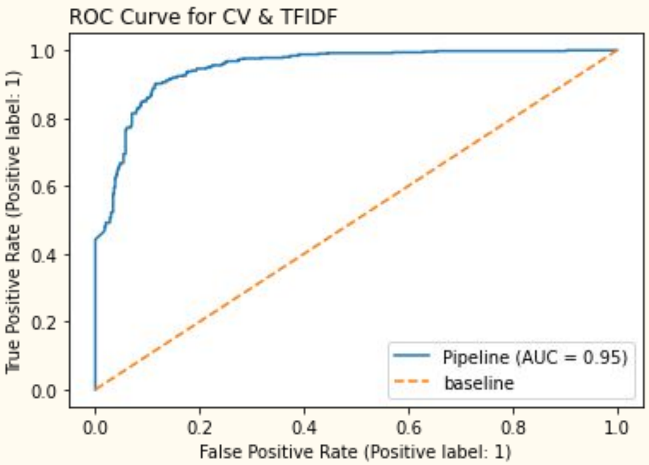




Random Forest using TFIDF

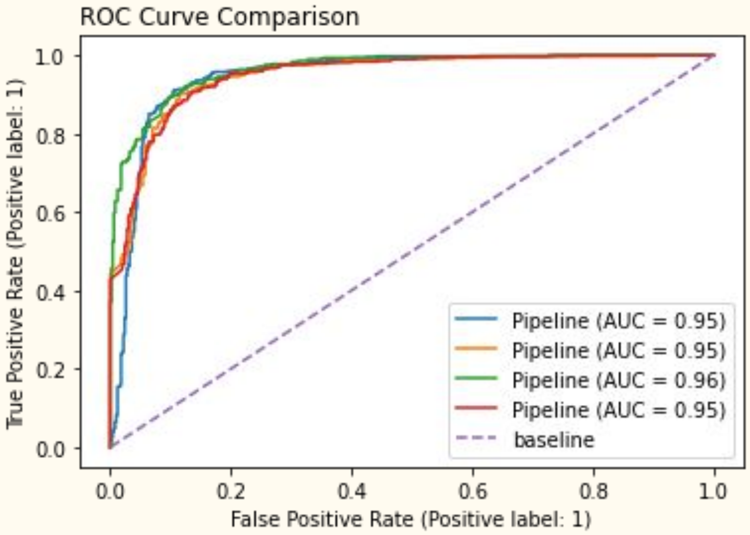
While this combination improved on the previous Random Forest in terms of false negatives, the accuracy score was still about the same for test data. With a much higher training data score, it was also the most over-fit model.

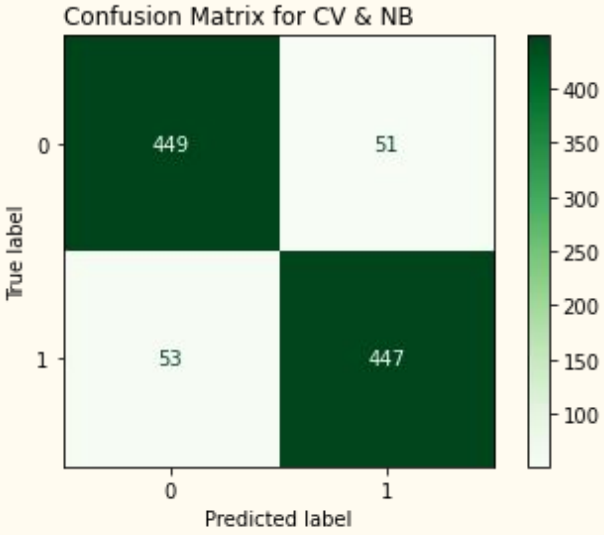




Winner: Naive Bayes & Count Vectorizer

While the accuracy of the best configuration (as determined through gridsearch) was roughly the same, the best balance of false negatives for predicting baking content was the model using Naive Bayes and Count Vectorizer.





Recommendations and Next Steps

- There appears to be evidence that these two cohorts have distinct needs and could be better served with differentiated content and marketing
- The winning model is a good starting point for further refinement. I'd also like to explore if the ~100 or so test data posts were simply indistinguishable "bayesian errors" or simply very short.

Thank You!

All images sourced from pixabay