Reddit Post Classification

• • •

Andrew Bergman
7 September 2019

Agenda

❖ Problem statement

❖ Subreddit background

❖ Preprocessing

❖ Modeling

❖ Evaluation

❖ Conclusions & Recommendations

❖ Sources

Problem Statement

❖ A data intern at Bon Appétit accidently deleted the `subreddit` tag from data they scraped from r/Cooking and r/AskCulinary . Their marketing team wants to target both subreddits for market and thus needs to differentiate the two. As a result, they approach us with the goal of predicting which posts came from r/Cooking.

Subreddit Background

- ❖ r/Cooking is a general, non-professional cooking community
  - ➢ It has ~1.35 million subscribers and is ranked 156th (per subredditstats.com)
- ❖ r/AskCulinary is an advice community, but also a place to share knowledge
 - ➢ It has ~235,000 subscribers and is ranked 1,002[nd] (per subredditstats.com)

Preprocessing

❖ Cleaning the data was fairly simple: we removed nulls, non-letters, cross-posts, & URLs

❖ We combined our two sets of texts: `title` and `selftext`

❖ The most common words from both subreddits were added to our stopwords
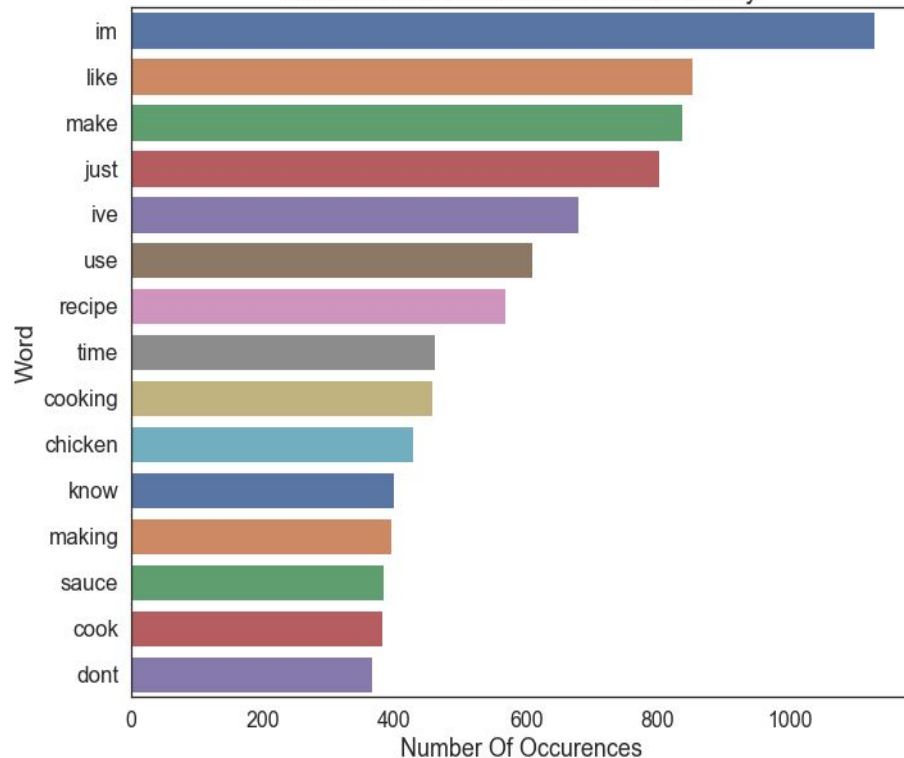
❖ The text was run through a lemmatizer

bon appétit

Most Frequent Words

Most Common Words From r/Cooking
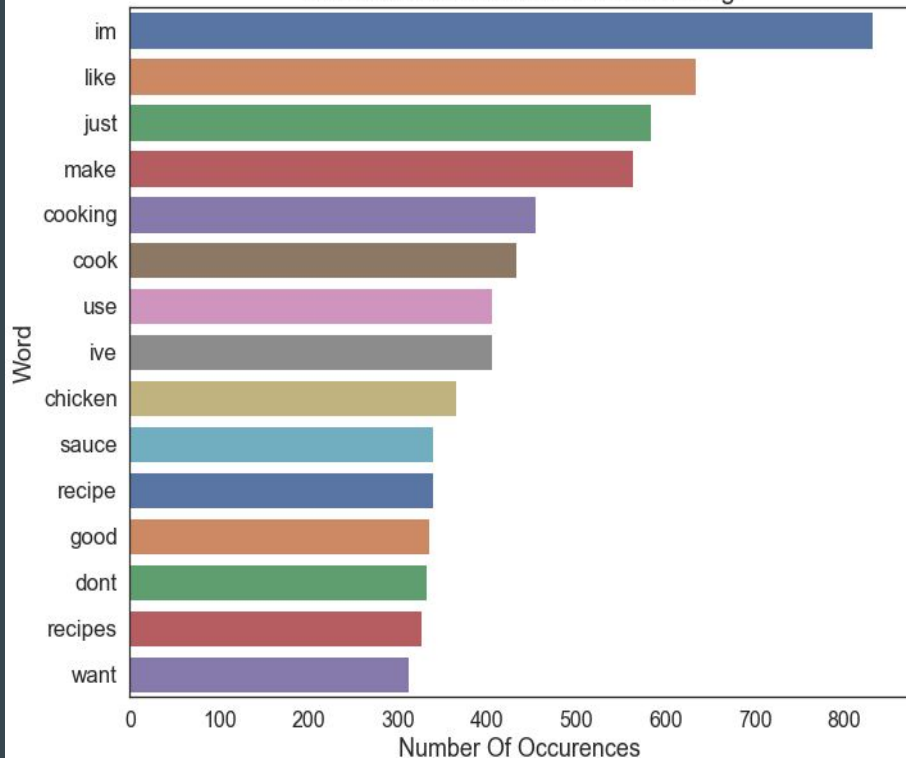
Most Common Words From r/AskCulinary

Modeling

❖ We used four models: logistic regression, SVC, random forest, & XGBoost
❖ Each model was run twice: once with count vectorization and once with TFIDF vectorization
❖ Every model optimized with gridsearching
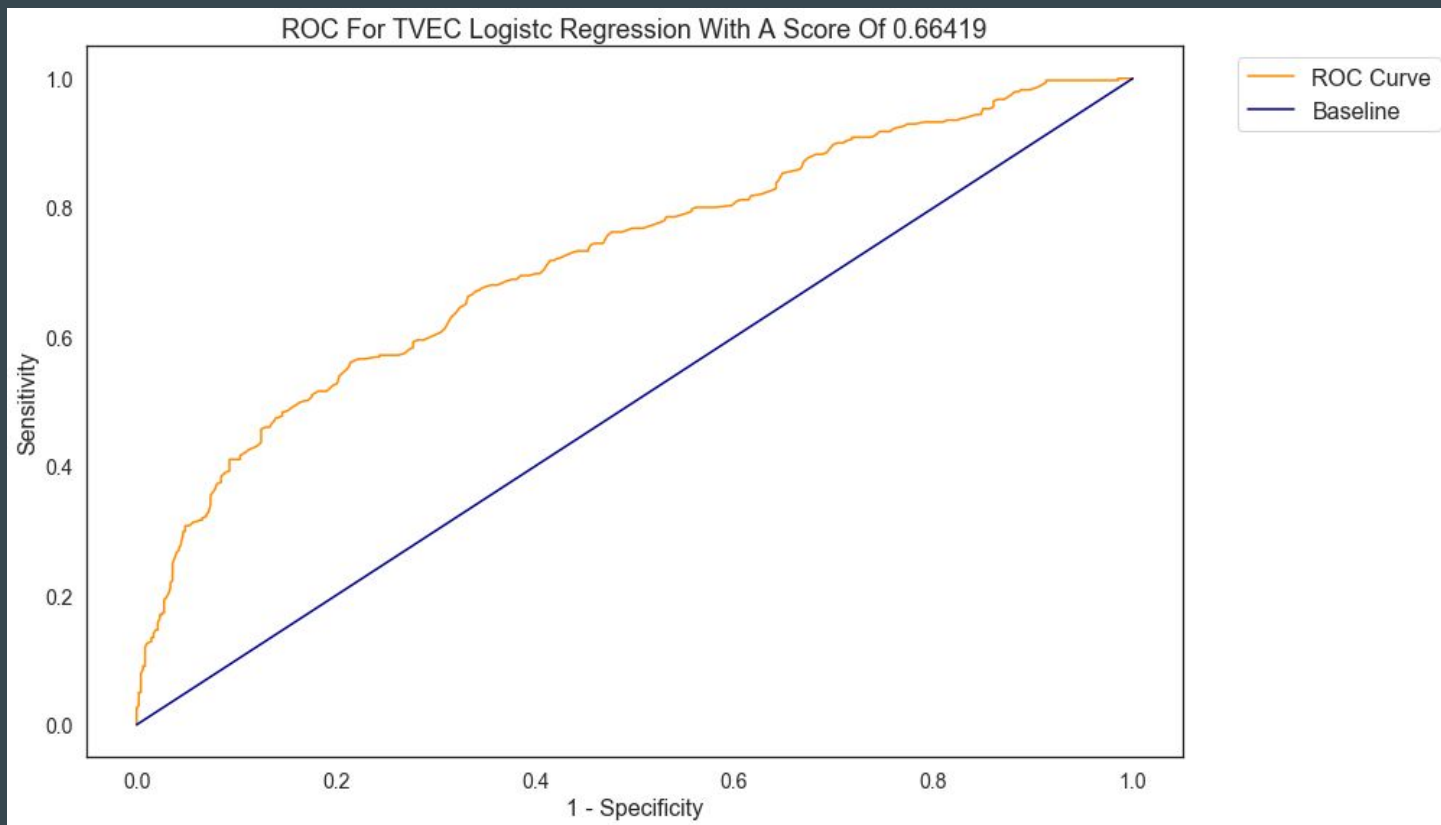❖ We gridsearched hyperparameters for the models and vectorizers

Evaluation

❖ The best model was a logistic regression with TFIDF vectorization
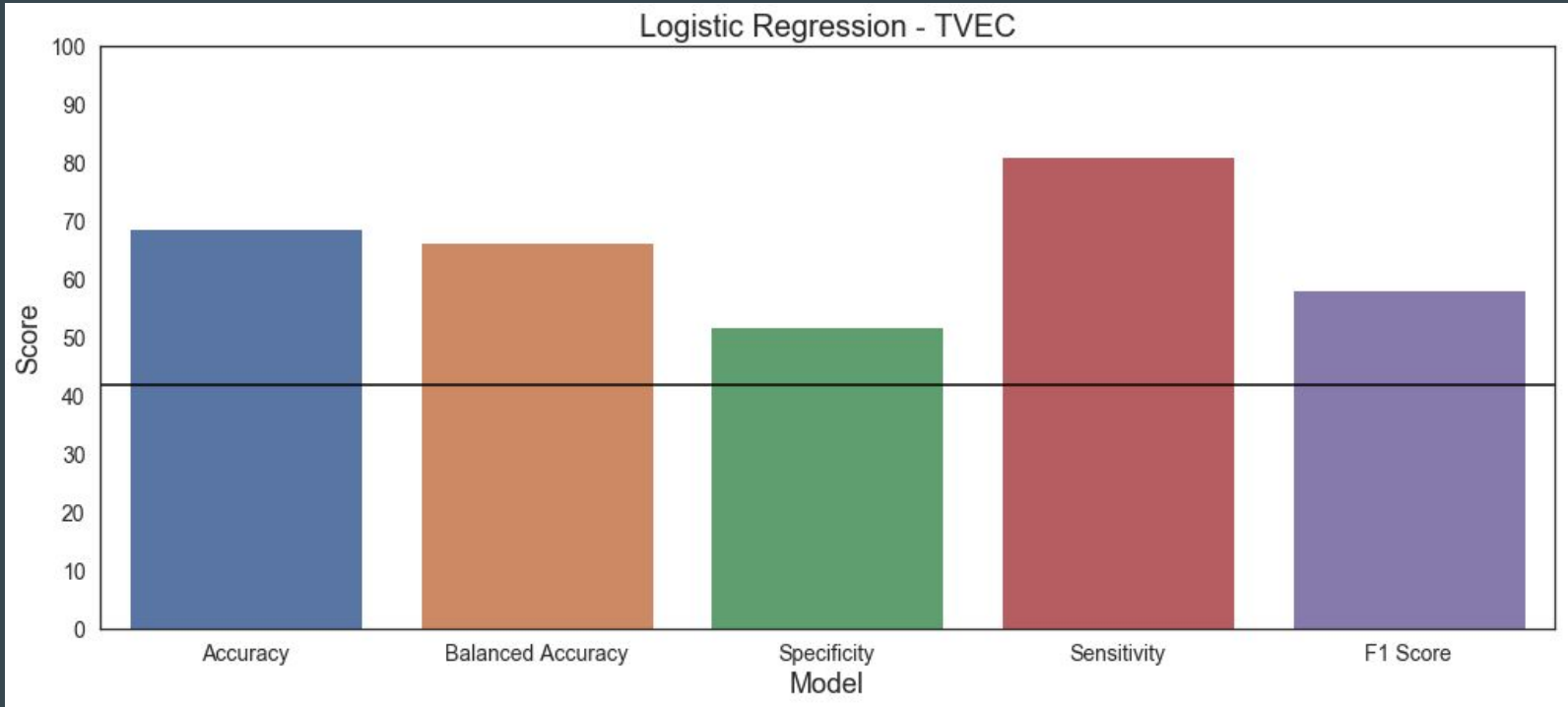
Evaluation

Conclusions & Recommendations

❖ We cannot recommend using our best model for distinguishing subreddits
❖ Our specificity was high but the sensitivity was low
❖ The model has a highly variable performance
❖ Going forward we want to experiment with different vectorizers
❖ We would also like to try more complex classifiers such as FFNNs

Sources

- ❖ https://subredditstats.com/r/Cooking
- ❖ https://subredditstats.com/r/AskCulinary