# Sentiment Analysis
# on amazon Fine Food Reviews

Darren Chia

# Executive Summary

## *Business problem*

To find an efficient way for product owners to gather feedback on their products

## *Use case*

Product owners can react to the feedback

# Executive Summary

## *Solution*

To identify keywords relating to positive and negative reviews via classification algorithm

# Dataset

Source: *kaggle*

568,000 reviews from *Amazon*

256,000 users

74,000 food products

13 years from Oct 1999

# Dataset

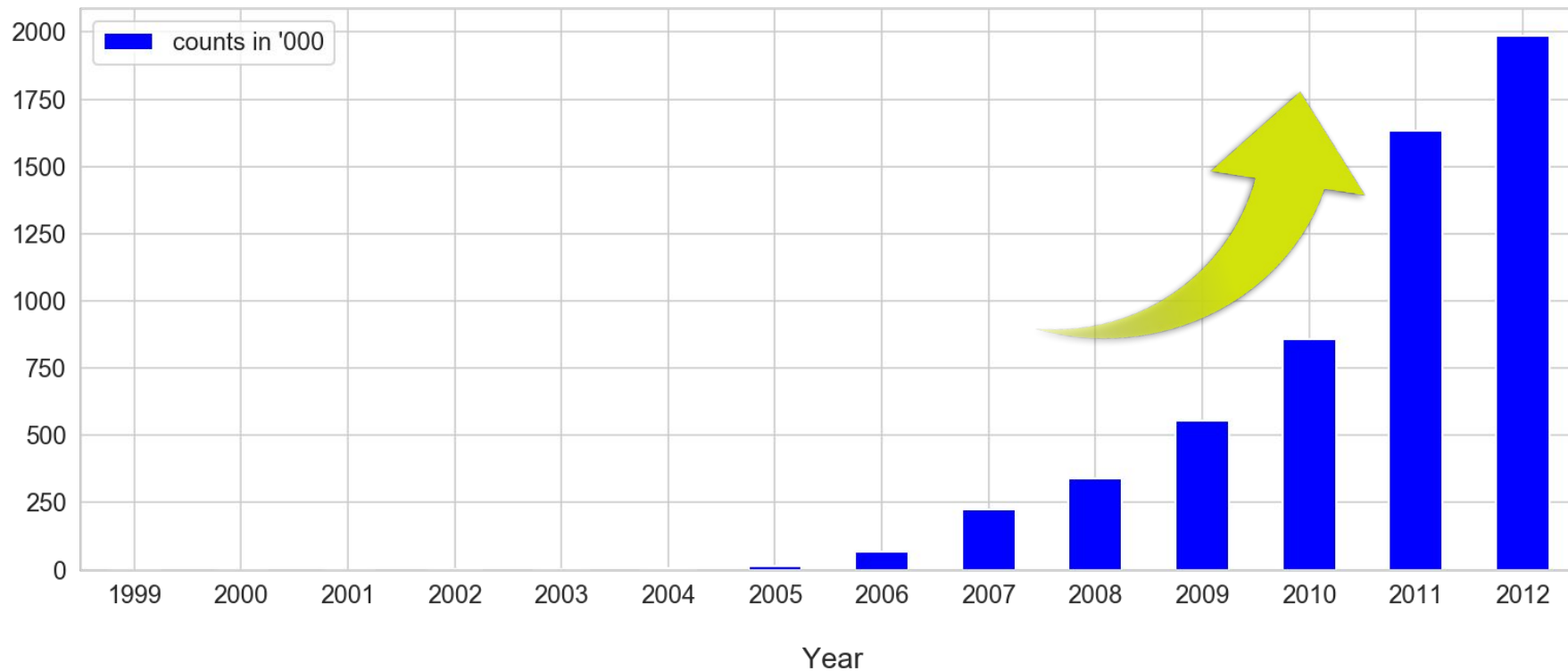| User ID | Product ID | Review |
|---|---|---|
| Name | | Review Time |
| Helpful Votes | | Summary |
| Total Votes | | |
| **Score** | | |

# Data Cleaning

🕐 converted UTC time:

'1303862400' to '2011-04-27 08:00:00'

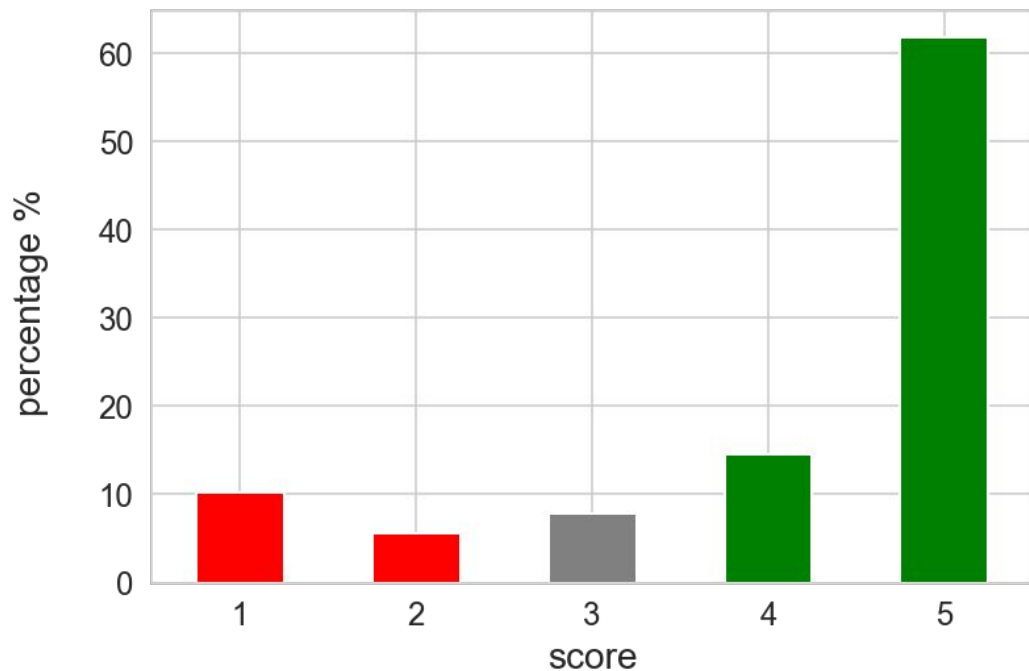🗑 removed duplicated reviews from the same user
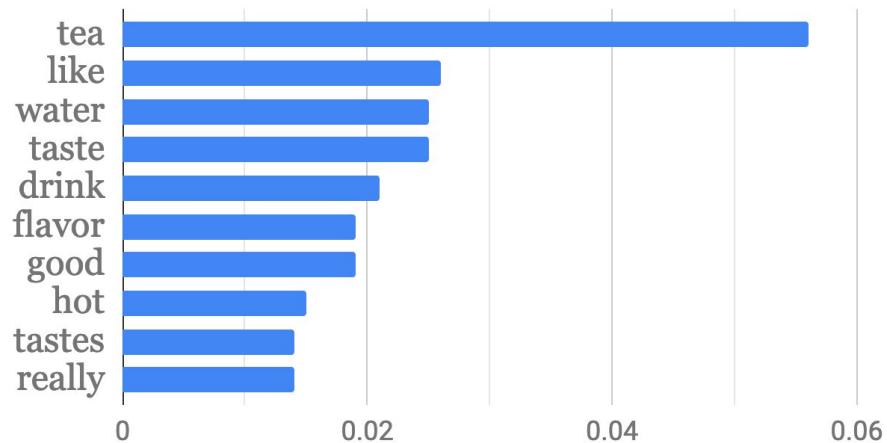
# EDA

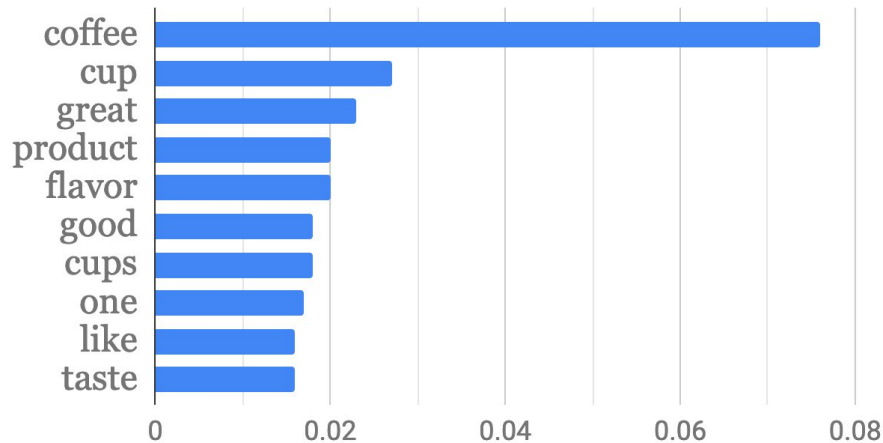## Reviews over time

# EDA



percentage of reviews for each rating

- Imbalanced classes

- 50% of 2012 reviews
  - **~ 91,500**
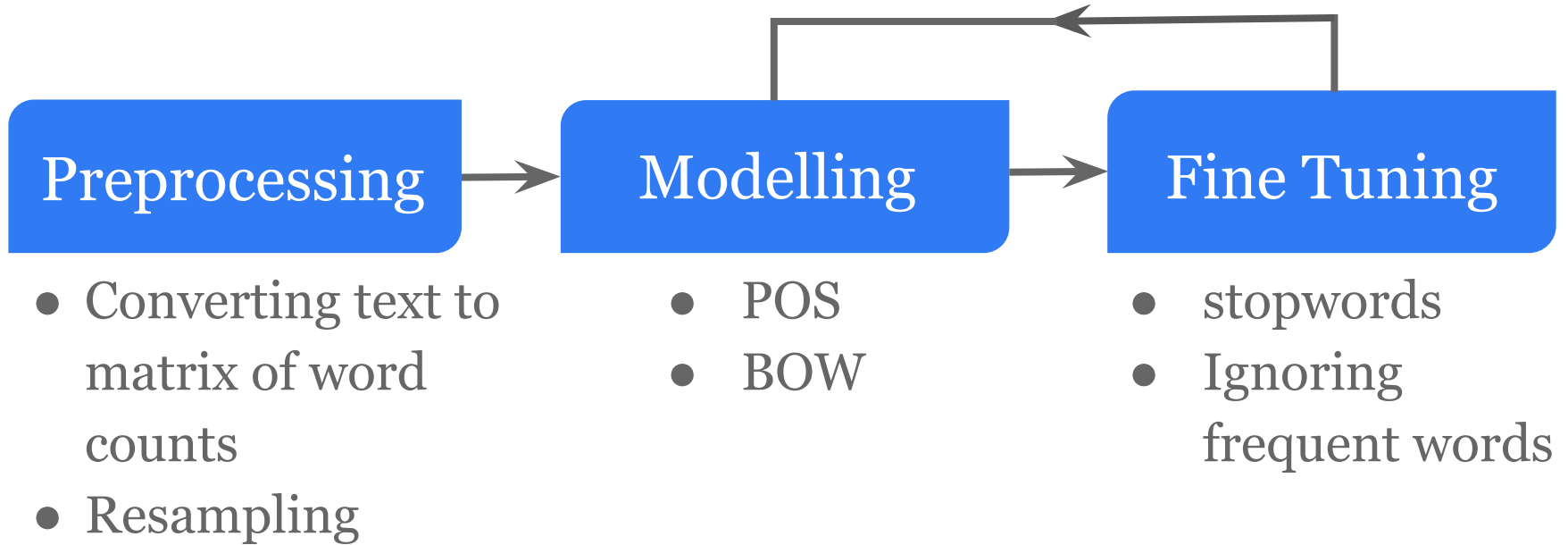
# LDA Topic Modelling

**Topic 1: Tea**

| | |
|---|---|
| tea | ████████████████████████ |
| like | ███████████ |
| water | ██████████ |
| taste | ██████████ |
| drink | ████████ |
| flavor | ███████ |
| good | ███████ |
| hot | ██████ |
| tastes | █████ |
| really | █████ |

0    0.02    0.04    0.06

**Topic 2: Coffee**

| | |
|---|---|
| coffee | ██████████████████████████ |
| cup | █████████ |
| great | ████████ |
| product | ███████ |
| flavor | ███████ |
| good | ██████ |
| cups | ██████ |
| one | ██████ |
| like | █████ |
| taste | █████ |

0    0.02    0.04    0.06    0.08

Other topics: Snacks, Cat food, Dog food

# Overall Approach

| Preprocessing | Modelling | Fine Tuning |

- Converting text to matrix of word counts
- Resampling

- POS
- BOW

- stopwords
- Ignoring frequent words

# Selected Models

| Multinomial Naive Bayes Classifier | Logistic Regression Classifier with Regularization | Random Forest Classifier |
|---|---|---|
| Predictions based on probabilities | | Aggregate predictions |
| Able to interpret top features | | |

# Evaluation Metrics



1. **Area under the ROC curve (AUC)**:
   ○ shows how well a model can distinguish classes
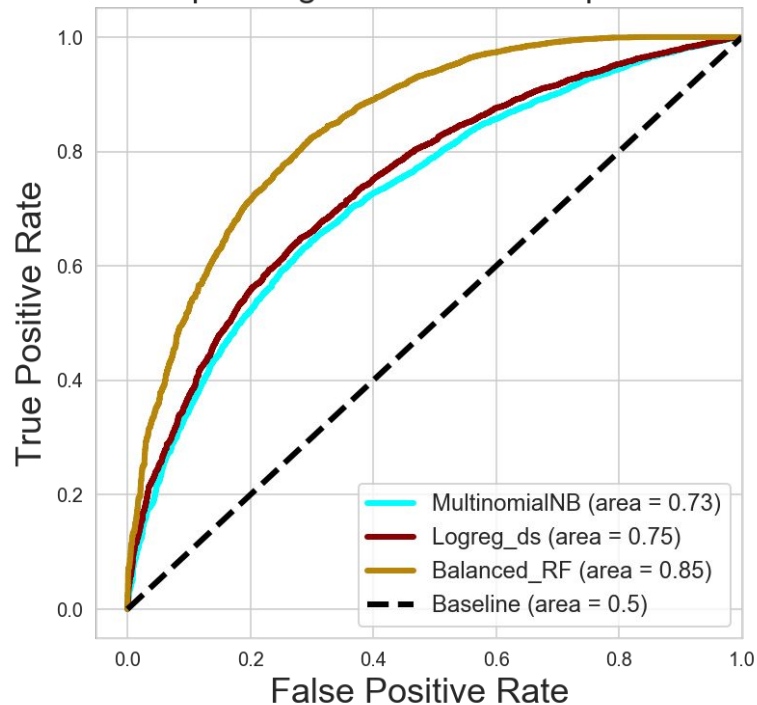
2. **Precision**:
   ○ shows for the predictions, how many are indeed in the right class
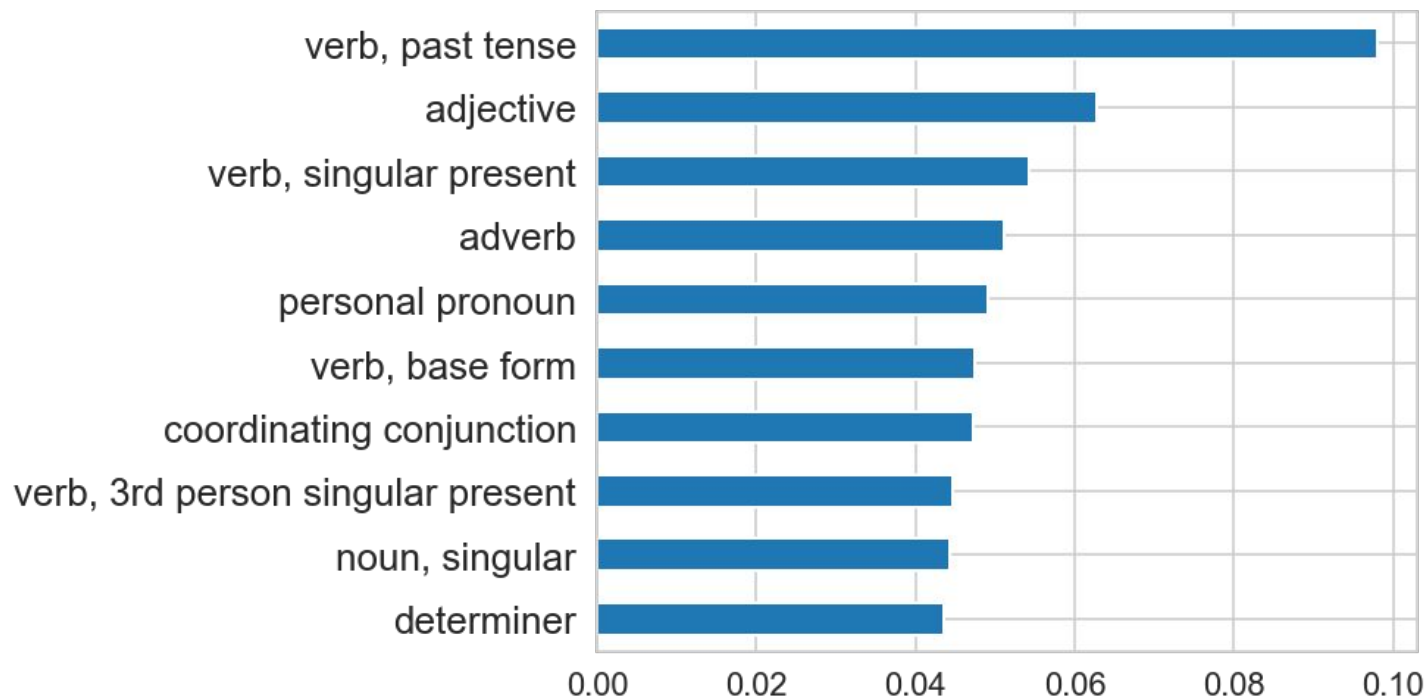
**Higher = Better**

# POS Classification

# Findings

Receiver operating characteristic: is positive review



| Model | AUC | Class 1 | |
|---|---|---|---|
| | | Recall | Precision |
| Naive Bayes MultinomialNB | 0.73 | 0.67 | 0.91 |
| **Balanced Random Forest Classifier** | **0.85** | **0.77** | **0.94** |
| Logistic Regression + Regularization | 0.75 | 0.56 | 0.93 |

# Findings



Feature Importance

| | |
|---|---|
| verb, past tense | |
| adjective | |
| verb, singular present | |
| adverb | |
| personal pronoun | |
| verb, base form | |
| coordinating conjunction | |
| verb, 3rd person singular present | |
| noun, singular | |
| determiner | |

x-axis: 0.00, 0.02, 0.04, 0.06, 0.08, 0.10

# Findings

| POS |
| --- |
| verb, past tense |
| adjective |
| verb, singular present |
| adverb |
| personal pronoun |
| verb, base form |

*"**Great** product, **nice** combination of chocolates and **perfect** size!  The bags had plenty, and they were shipped **promptly**. The kids in the neighborhood **liked** our candies!"*

👍

*"This Wolfgang Puck Coffee Jamaica Me Crazy is the **worst** coffee I've ever had! I **threw** it out right away after one sip! It taste like all coconut flavoring. **Hard** to believe someone thought this taste good and was marketable. They should be **fired**!"*
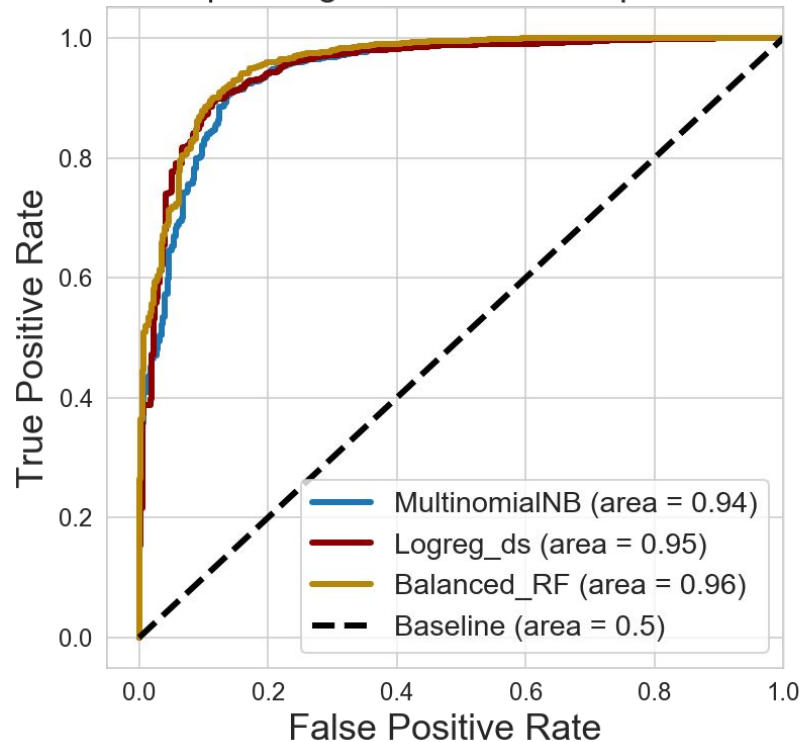
👎

# BOW Classification

# Coffee Reviews 2012
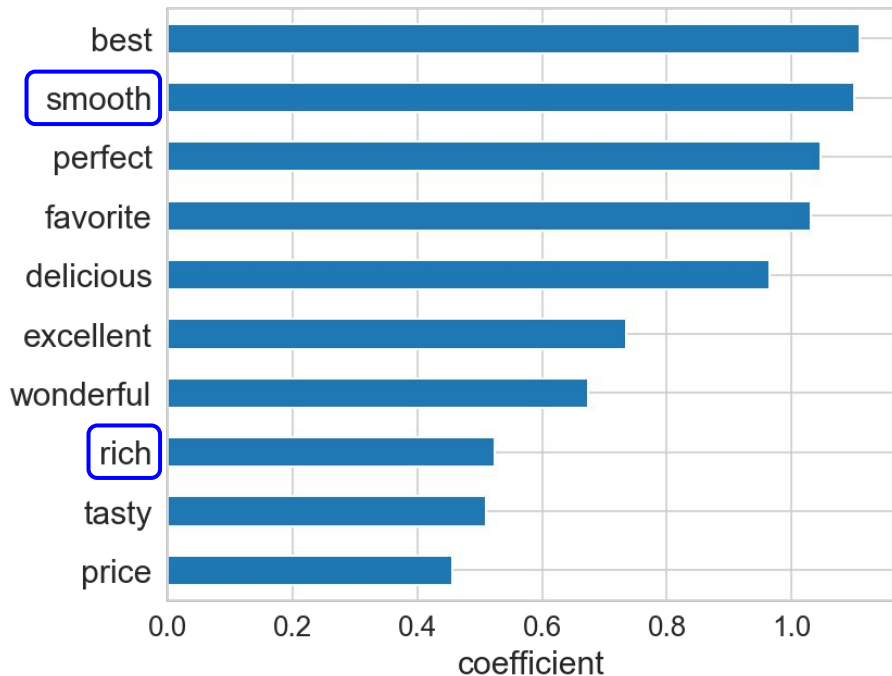
# Findings

Receiver operating characteristic: is positive review



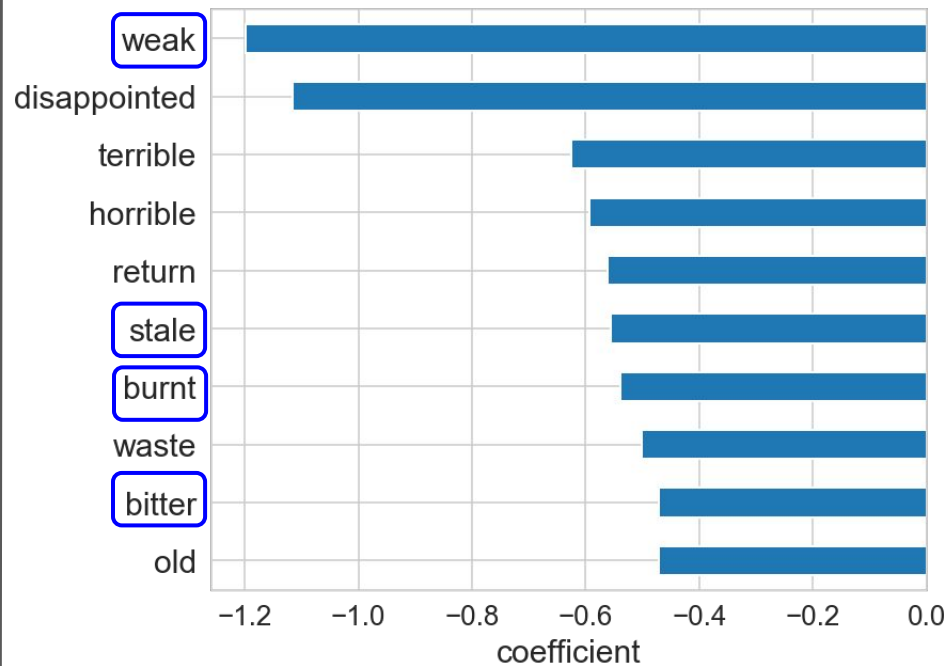| Model | AUC | Class 1 | |
|---|---|---|---|
| | | Recall | Precision |
| Naive Bayes MultinomialNB | 0.94 | 0.90 | 0.97 |
| Balanced Random Forest Classifier | 0.96 | 0.94 | 0.97 |
| **Logistic Regression + Regularization** | **0.95** | **0.89** | **0.98** |

# Findings



Positive Keywords

Negative Keywords

# Limitations

- There may be **fake reviews**
- Product names and categories are not provided

# Next step

To gather feedback on other product categories
e.g. pricing, quality