Coffee & Tea: reddit analysis

Luke Heeringa

Overview

- Problem Statement
- Methodology
- Model Performance & Comparison
- Key Findings & Examples
- Conclusions

Classification: Coffee vs Tea

DATA COLLECTION

5,000 posts each from r/Coffee and r/tea

November 22, 2020 - January 13, 2021

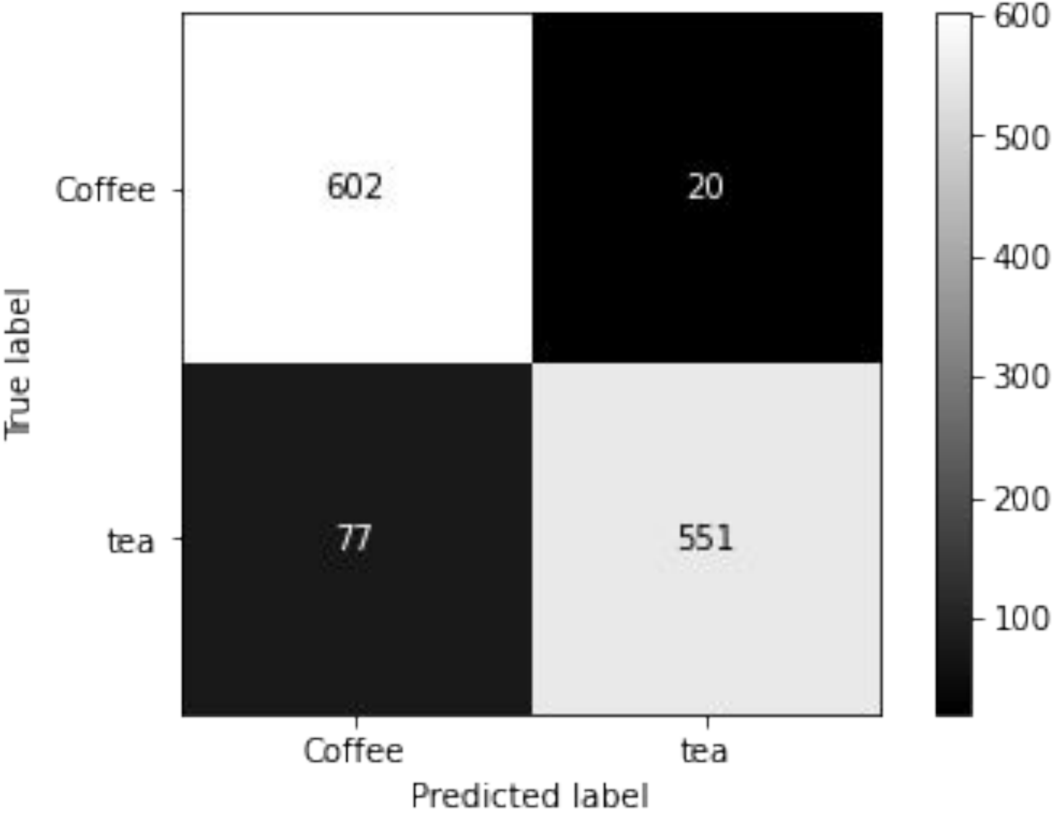
Model Performance

Evaluated on Accuracy

Train/Test Split = 75/25

	Train Data	Test Data
Logistic Regression	96.2%	92.2%
Random Forest	95.5%	91.4%
Support Vector	94.5%	90.1%
Naive Bayes	94.6%	89.5%
Null	50.0%	50.0%

Confusion Matrix



Most Significant Features

r/Coffee

coffee grinder

espresso v60

moka beans

aeropress pour

machine moccamaster

r/tea

tea teas

matcha teapot

oolong chai

green gaiwan

leaf sencha

High Probability Examples

“ COFFEE CUPS, COFFEE TYPE FOR COFFEE LOVERS. Espresso, Americano, Frappe, Cappuccino, AND Mocha ”

P(r/Coffee) > 99.99%

“ Tea (Camellia Sinensis) The tea in Sri Lanka is so special and known for its high-quality factor. Ceylon tea, as it has been known since the 19th century, has been the base tea of choice for most tea customers around the world. ”

P(r/tea) > 99.99%

Low Probability Examples

“ Single origin vs blends? ”

P(r/Coffee) = 50.11%

P(r/tea) = 49.89%

Actual: r/tea

“ I have the weirdest feeling that chamomille and pineapple go together really well ”

P(r/Coffee) = 50.05%

P(r/tea) = 49.95%

Actual: r/tea

Conclusions:

- Ability to identify coffee and tea consumers
- Market opportunity for tea branding
- Model improvement via robust word recognition

Question & Answer