Web APIs & NLP

Predicting between the coffee and tea subreddits

Problem Statement

- Up and coming coffee & tea company
- Leverage Data to form sales strategy
- Aid of classification models

Background

COFFEE AND TEA (2/3 MOST POPULAR BEVERAGES WORLDWIDE

TWO OF THE MOST PROFITABLE COMMODITIES.

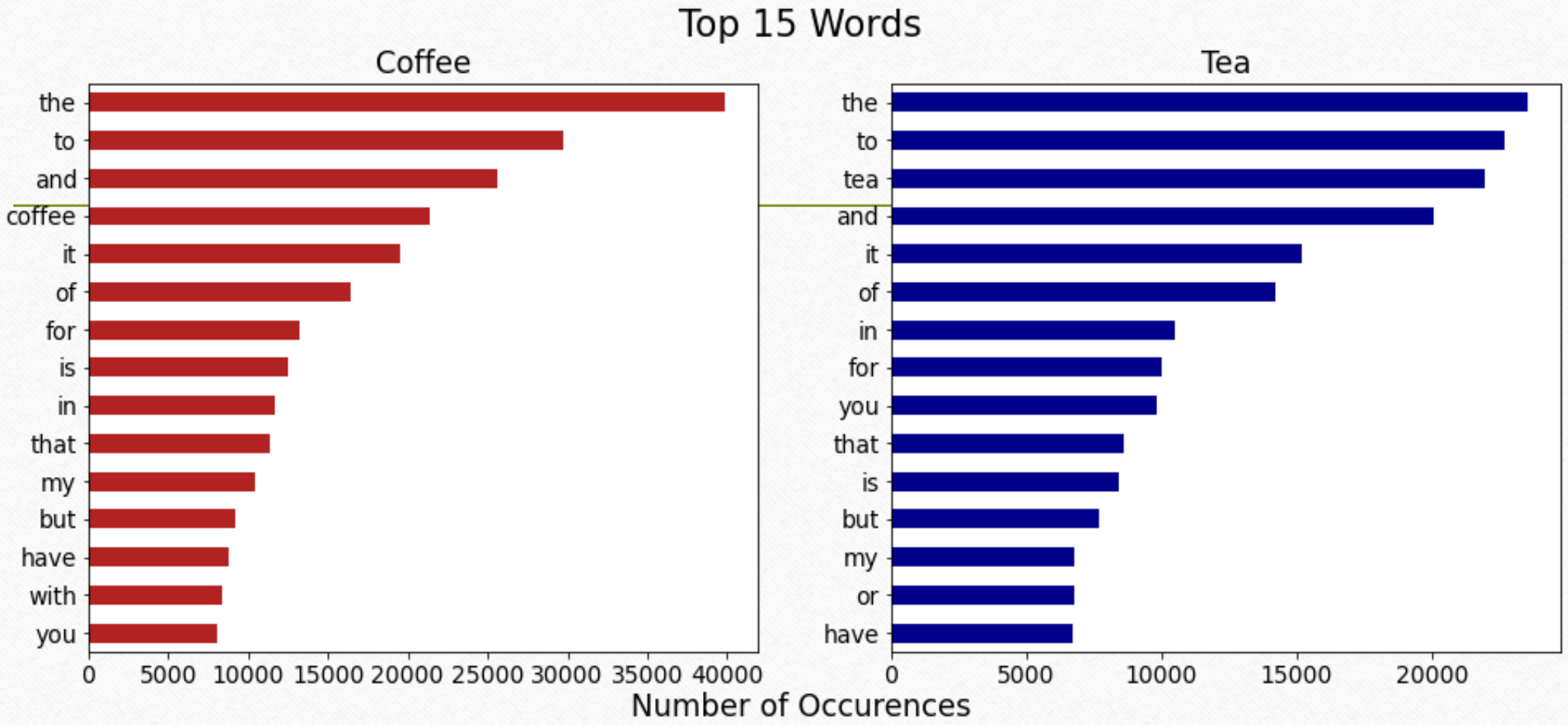
PEOPLE LOVE CAFFEINE

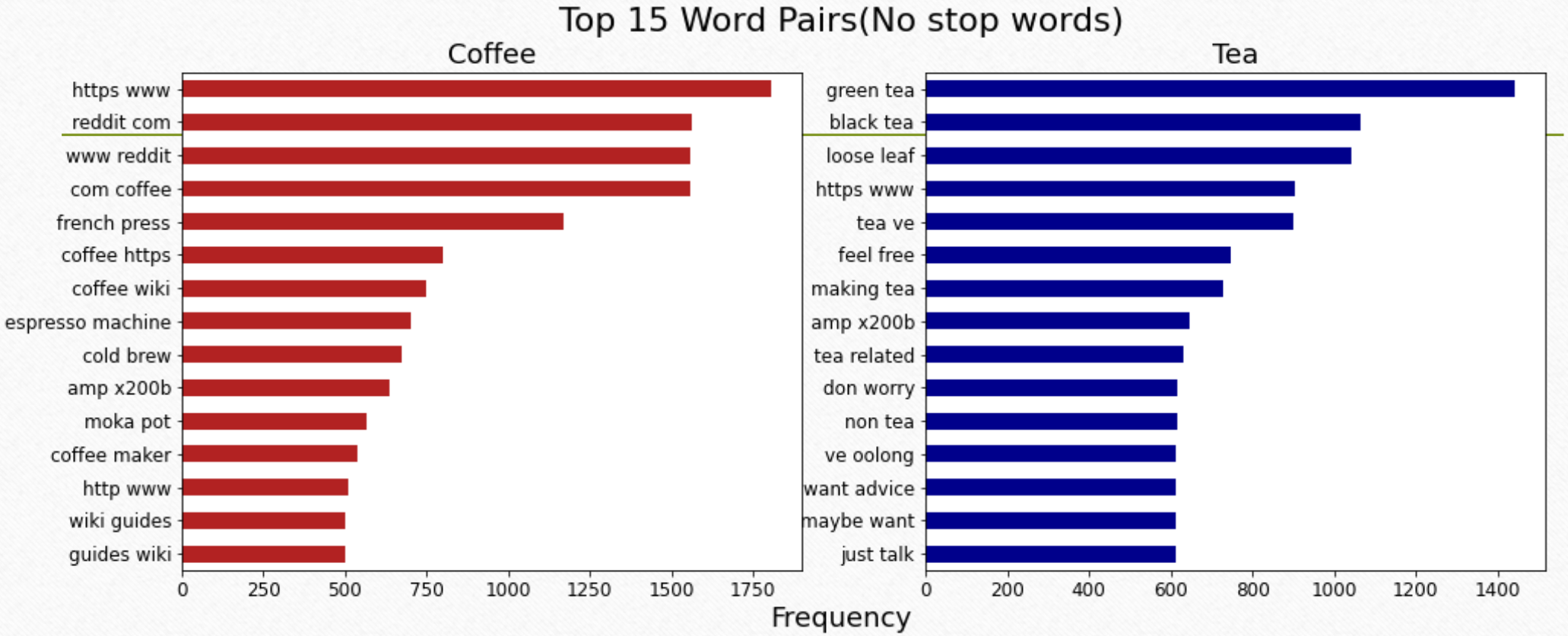
The Data

PUSHSHIFT API

SCRAPED TEA &
COFFEE SUBREDDITS

19592 POSTS


Top 15 Words
Coffee
the
to
and
coffee
it
of
for
is
in
that
my
but
have
with
you
0
5000
10000
15000
20000
25000
30000
35000
40000
Tea
the
to
tea
and
it
of
in
for
you
that
is
but
my
or
have
0
5000
10000
15000
20000
Number of Occurences


Top 15 Word Pairs(No stop words)
Coffee
https www
reddit com
www reddit
com coffee
french press
coffee https
coffee wiki
espresso machine
cold brew
amp x200b
moka pot
coffee maker
http www
wiki guides
guides wiki
0
250
500
750
1000
1250
1500
1750
Tea
green tea
black tea
loose leaf
https www
tea ve
feel free
making tea
amp x200b
tea related
don worry
non tea
ve oolong
want advice
maybe want
just talk
0
200
400
600
800
1000
1200
1400
Frequency

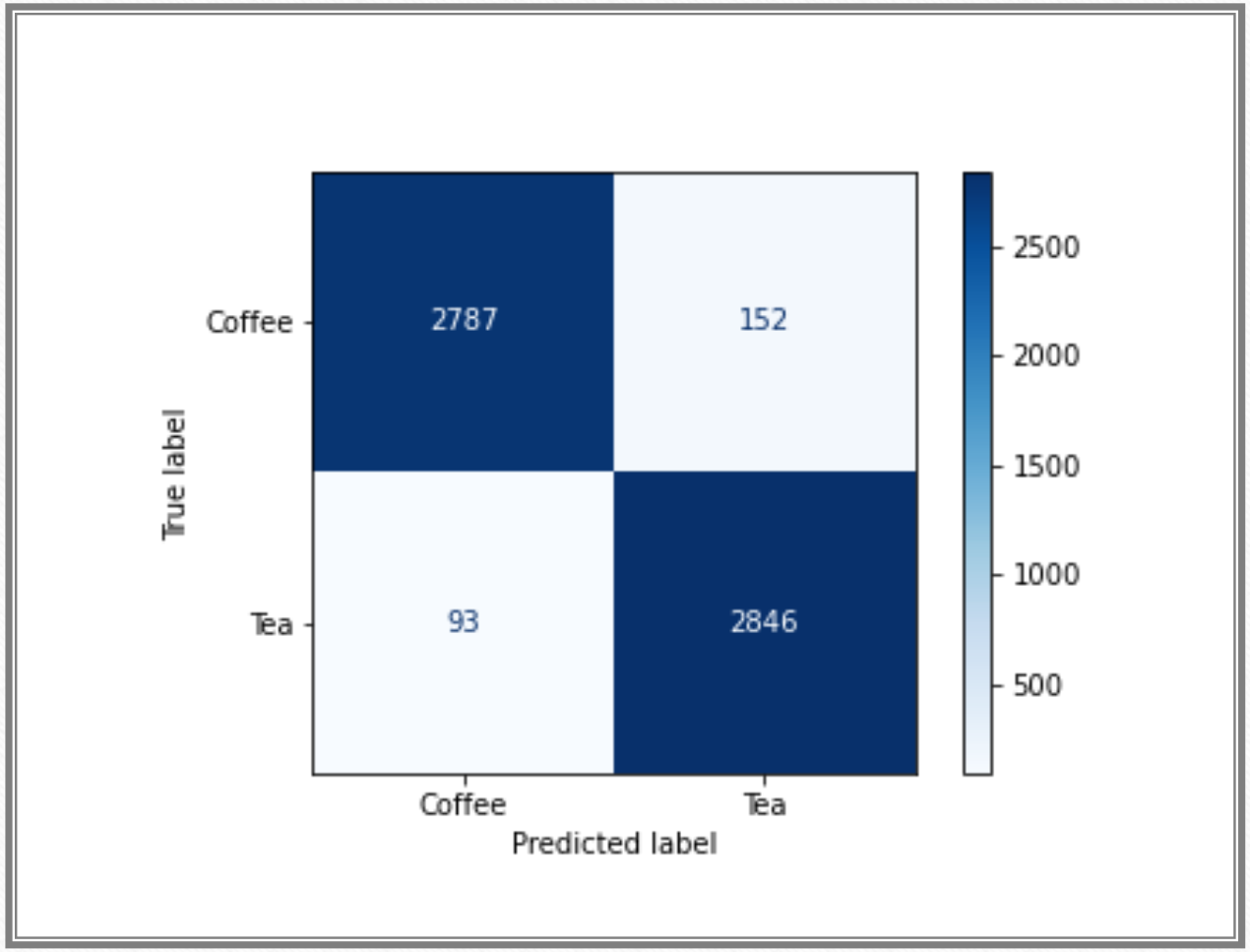
Model Evaluation

ACCURACY

RECALL

SPECIFICITY

TYPE I AND TYPE II

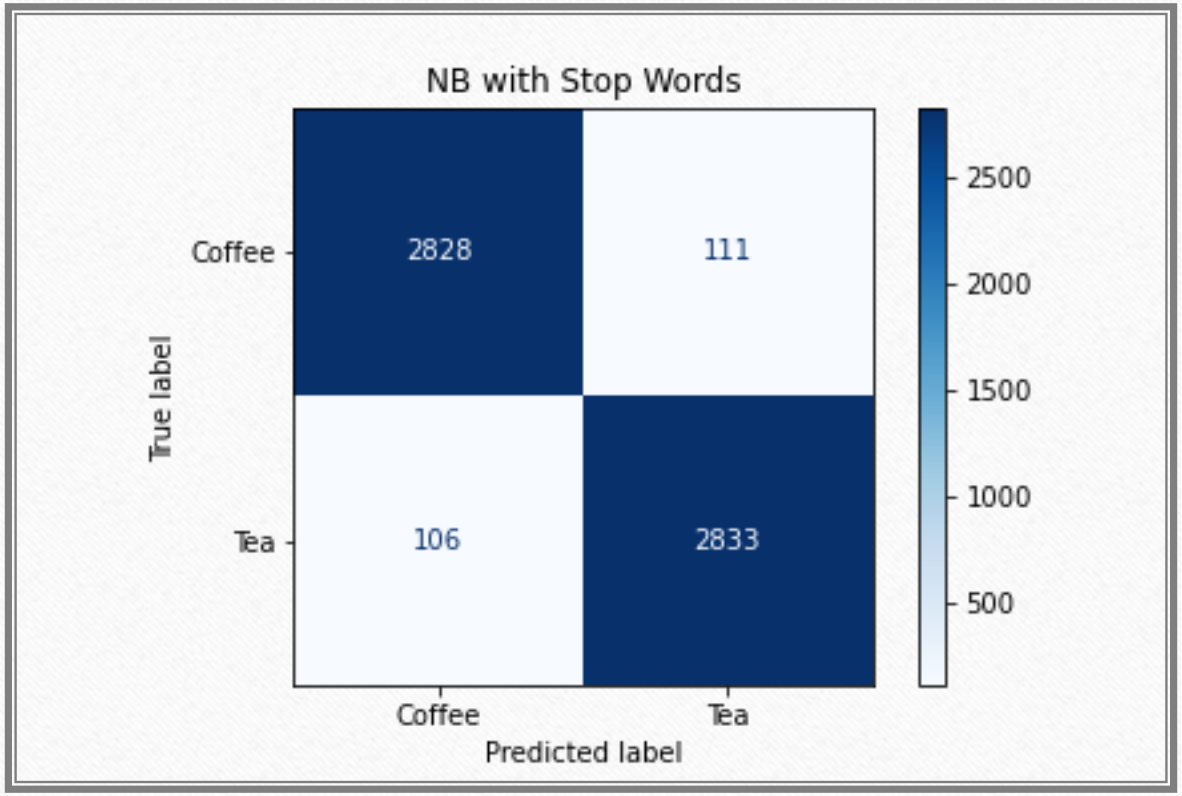


Multinomial NB

	Multinomial NB Model
Training Score	96.1%
Testing Score	95.8%
Specificity	94.8%
Recall	96.8%
Type I error	5.2%
Type II error	3.2%

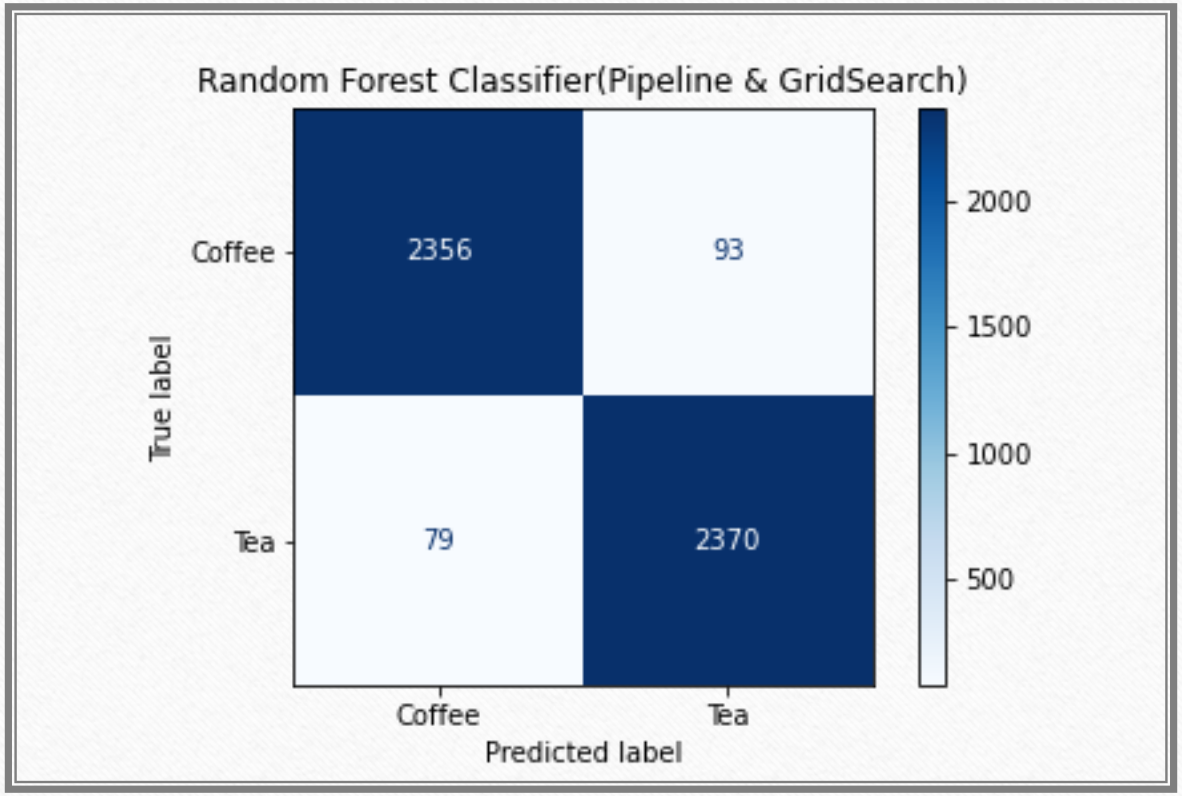
Multinomial NB(w/stop words)

	Multinomial NB (w/Stop words)
Training Score	96.2%
Testing Score	96.3%
Specificity	95.4%
Recall	96.4%
Type I error	3.8%
Type II error	3.6%



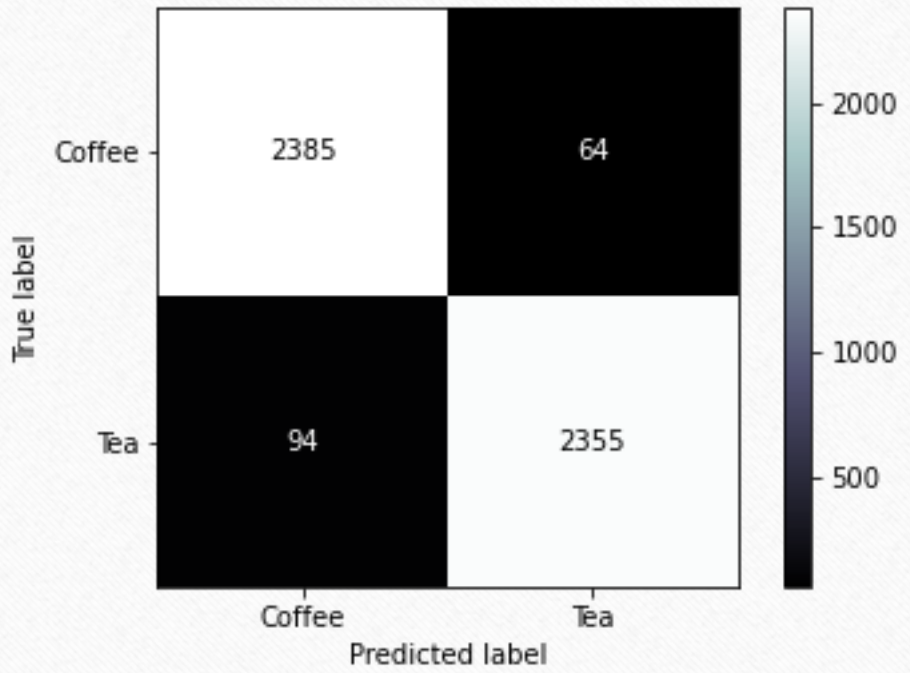
Random Forest Classification Model

	Random Forest Classifier (Pipeline& Grid search)
Training Score	99.9%
Testing Score	96.5%
Specificity	96.1%
Recall	98.3%
Type I error	3.8%
Type II error	3.2%



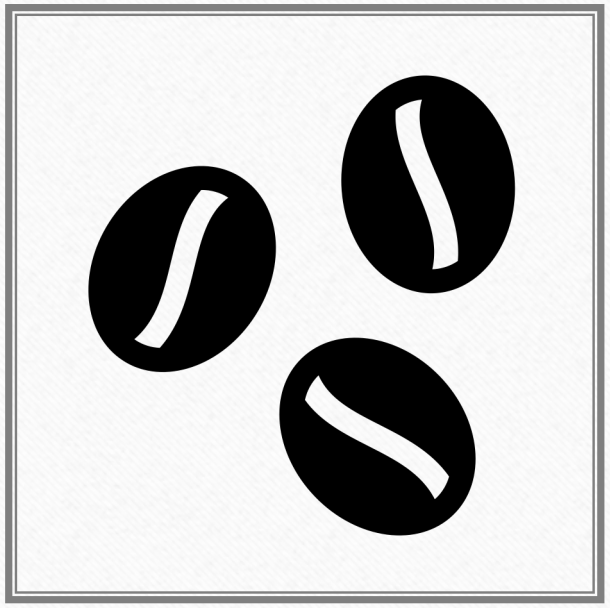
Logistic Regression

	Logistic Regression(TFID/ GS)
Training Score	99.9%
Testing Score	96.8%
Specificity	97.4%
Recall	96.2%
Type I error	2.6%
Type 2 error	3.8%



	Multinomial NB Model	Multinomial NB (w/Stop words)	Random Forest(TFID/ GS)	Random Forest Classifier (Pipeline& Grid search)	Logistic Regression(TFID)
Training Score	96.1%	96.2%	99.9%	99.9%	98.9%
Testing Score	95.8%	96.3%	96.3%	96.5%	96.8%
Specificity	94.8%	95.4%	96.1%	96.1%	97.4%
Recall	96.8%	96.4%	96.5%	98.3%	96.2%
Type I error	5.2%	3.8%	3.9%	3.8%	2.6%
Type II error	3.2%	3.6%	3.5%	3.2%	3.8%

Conclusion



- Word Pairs tell a story
- Logistic Regression performed the best
- Looking at a company's sales data to complete the picture