PERSONAL FINANCE
OR
FINANCIAL INDEPENDENCE

by Osei Anom

EXECUTIVE SUMMARY:

- collect unstructured data, pre-process that data, then build a predictive classification model.
- subreddit's "Personal Finance" and "Financial Independence”
- Which classification model best predicts the subreddit a post will fall under

PROCESS

1 Extracted Data using Pushshift API

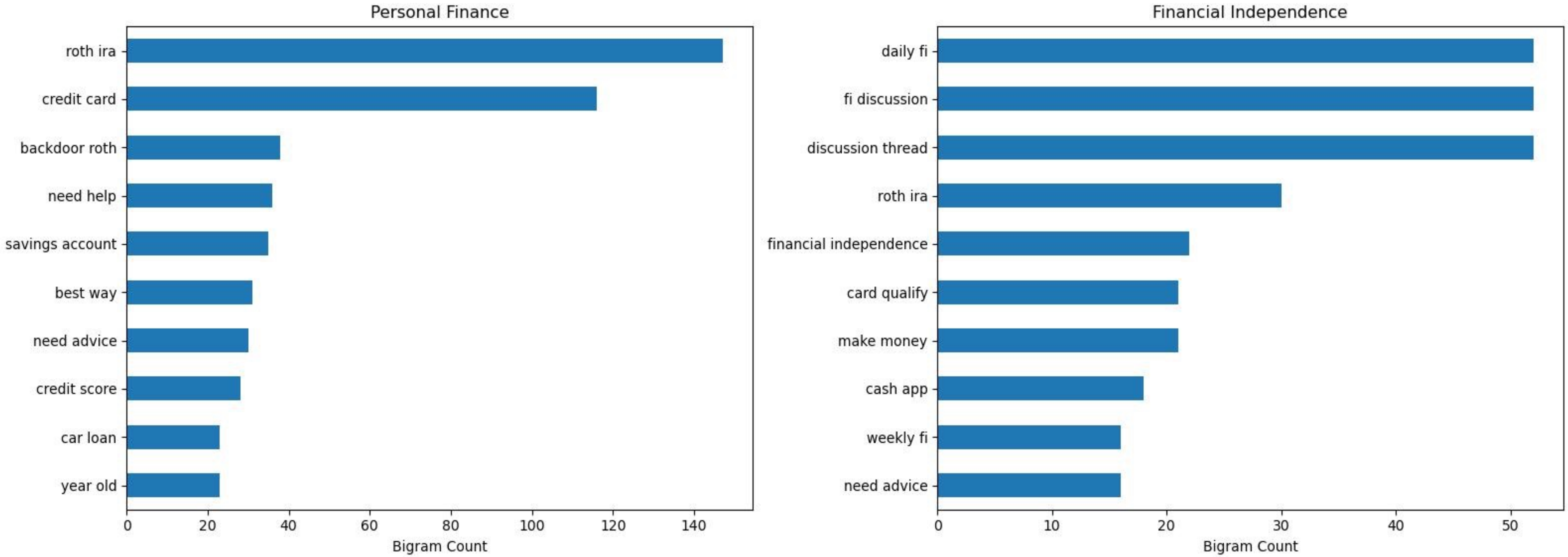
2 Exploratory Data Analysis

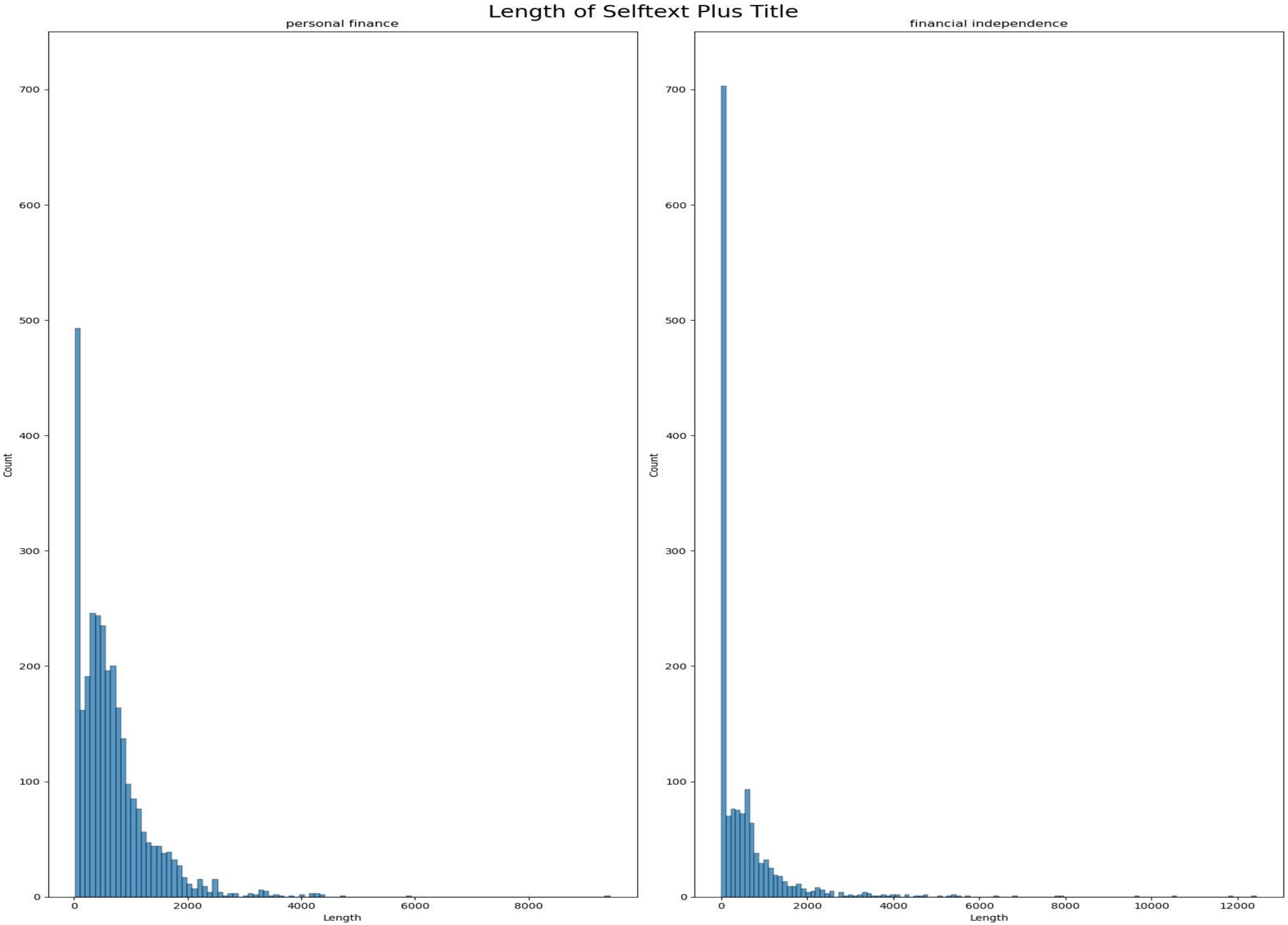
3 Initial Modeling

4 Secondary Modeling with additional data

* Adding selftext with title

Top Bigrams by Subreddit




Length of Selftext Plus Title
personal finance
financial independence
Count
Count
Length
Length
0
100
200
300
400
500
600
700
0
2000
4000
6000
8000
10000
12000

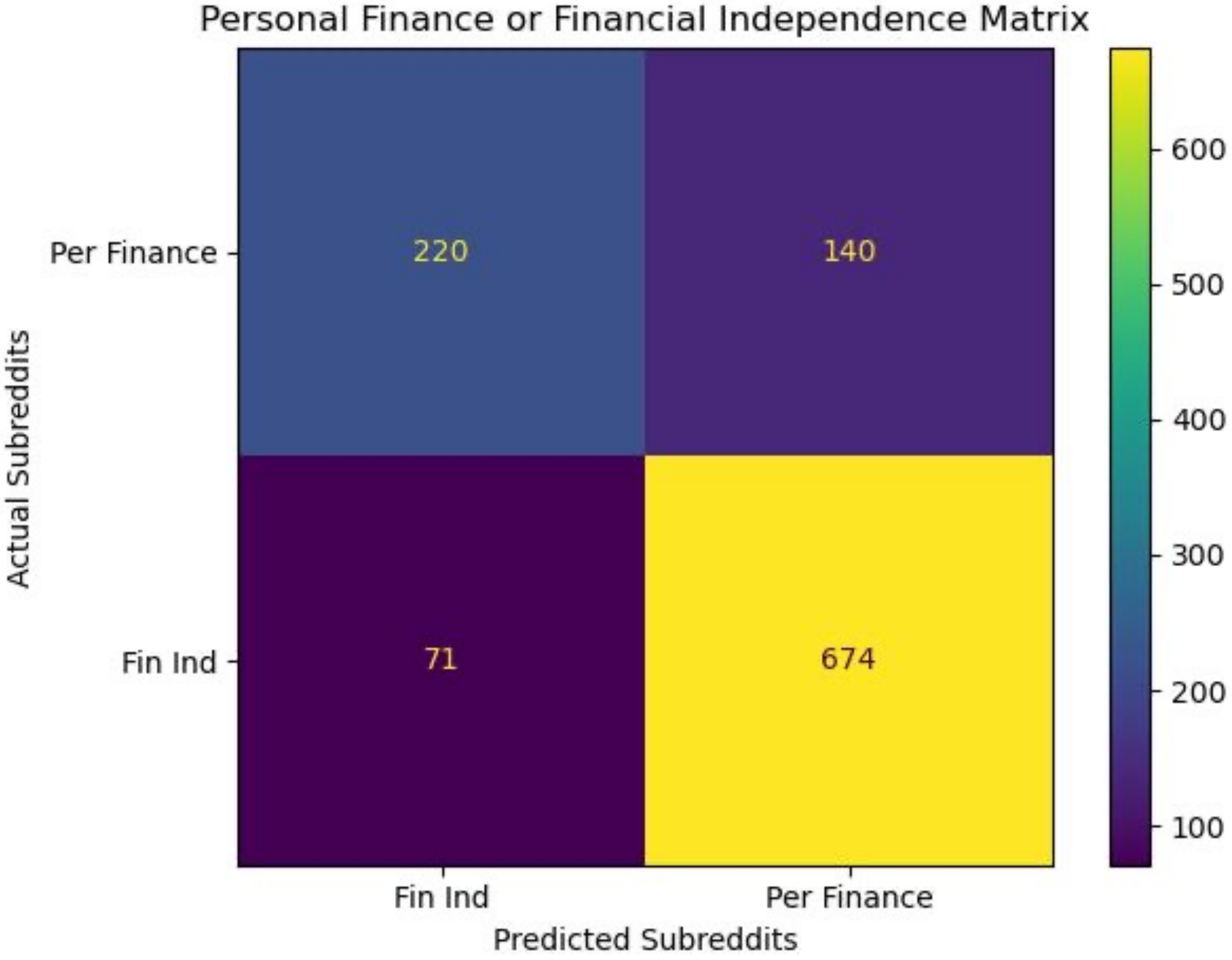
METHODOLOGY, INFERENCES, ASSUMPTIONS:

- Used Tfid Vectorizer through gridsearch running Logistic Regression, Decision Tree, Multinomial Naive Bayes, KNeighbors, Random Forest, AdaBoost, and finally Stacking
- Used F1 to determine performance between models
- Giving additional data would help increase test accuracy and F1 score

BEST PERFORMING MODELS (TEST SCORES)

	MN BAYES	RANDOM FOREST	ADA BOOST	STACKED
F1	84 %	85 %	84 %	86 %
Accuracy	76 %	78 %	78 %	81 %
Misclassification	24 %	22 %	22%	19 %

EVALUATION


Personal Finance or Financial Independence Matrix
Actual Subreddits
Per Finance
Fin Ind
220
140
71
674
Predicted Subreddits
Fin Ind
Per Finance
600
500
400
300
200
100

SUMMARY & CONCLUSIONS

- All models were overfit
- TBD if including word count and word length in post and self text would help.
- The model over predicted personal finance subreddit (false positives), partially due to the imbalance of classes
- Would like to explore over / under sampling, Grid Searching with balanced accuracy, synthetic data