Project 3:
Web APIs and Subreddit Classifier

Qi-Wen Ng

Outline

The Problem

Data Collection, Cleaning & Feature Engineering

Modelling & Evaluation

Conclusion & Recommendations

The Problem

Problem statement

Many people are now turning to social media as a source of information for personal financial advice. When they come with different financial goals, getting the right information would be dependent on the nature of the portals.

Problem statement

Many people are now turning to social media as a source of information for personal financial advice. When they come with different financial goals, getting the right information would be dependent on the nature of the portals.

How can we use predictive modelling to best predict which subreddit a post came from?

Who?

Individuals: Which subreddit should I obtain information, seek advice from or share my experience, given my financial situation?

Financial Advisors: What topics would my existing or potential clients be interested about that I should study into it?

Personal Finance Blogs / Websites: What topics are the community discussing about, and what kind of content can I share to improve engagements or visits?

The Process

Data Collection

Web Scrape

Reddit API

r/povertyfinance
306k members
992 posts

r/investing
1.0m members
1003 posts

Data Collection

Web Scrape

Cleaning & Preprocessing

Reddit API

r/povertyfinance
306k members
992 posts

r/investing
1.0m members
1003 posts

New:

- Column combining title and post content
- Adding binary class:
  - 1: investing
  - 0: poverty finance

Removed: duplicates, blanks, links, non-letters, stopwords

Tokenizing and Lemmatizing

Data Collection

Web Scrape

Reddit API

r/povertyfinance
306k members
992 posts

r/investing
1.0m members
1003 posts

Cleaning & Preprocessing

New:

- Column combining title and post content
- Adding binary class:
  - 1: investing
  - 0: poverty finance

Removed: duplicates, blanks, links, non-letters, stopwords

Tokenizing and Lemmatizing

Final Dataset

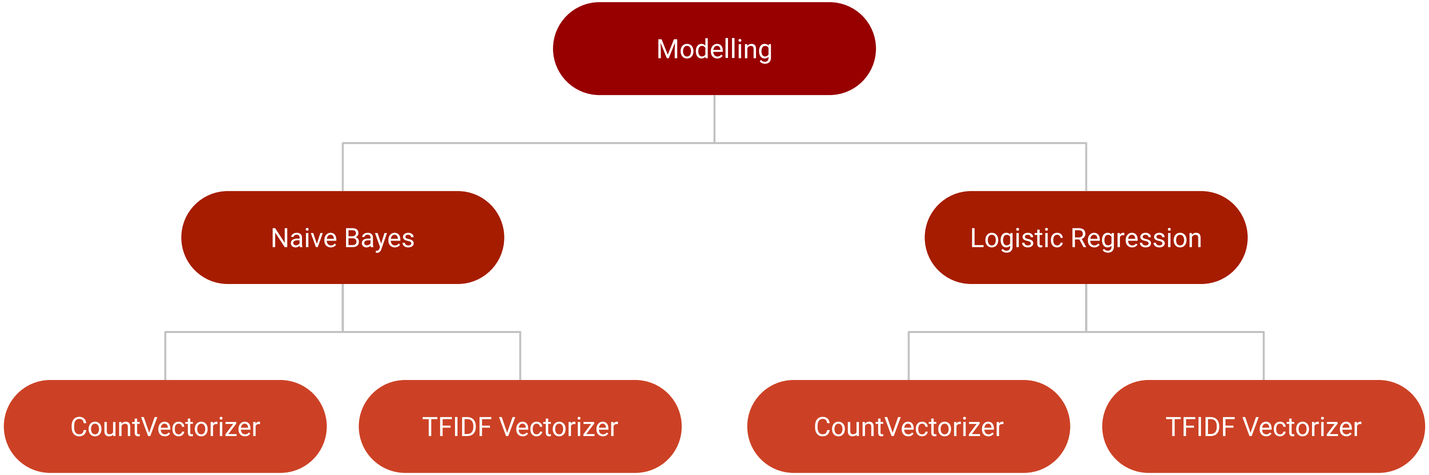
r/povertyfinance
992 posts

r/investing
898 posts

Split into Train, Validation & Testing Set

Modelling

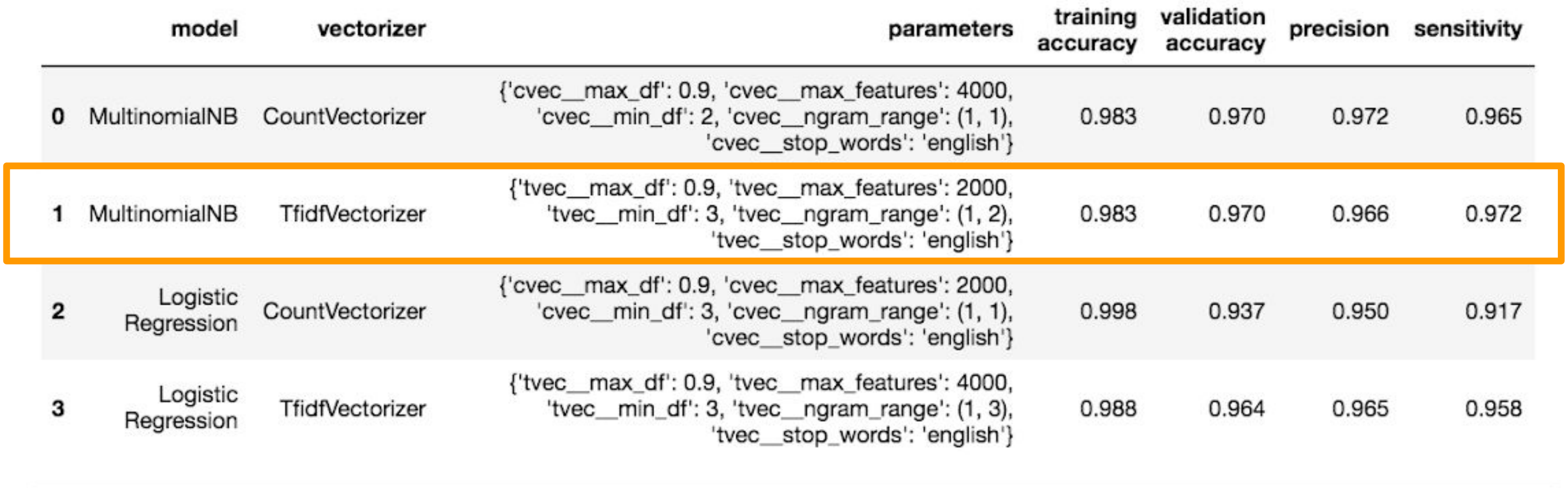
Modelling



Optimized using GridSearchCV

Model Metrics (Train & Validation)

Baseline proportion: 0.525



	model	vectorizer	parameters	training accuracy	validation accuracy	precision	sensitivity
0	MultinomialNB	CountVectorizer	{'cvec__max_df': 0.9, 'cvec__max_features': 4000, 'cvec__min_df': 2, 'cvec__ngram_range': (1, 1), 'cvec__stop_words': 'english'}	0.983	0.970	0.972	0.965
1	MultinomialNB	TfidfVectorizer	{'tvec__max_df': 0.9, 'tvec__max_features': 2000, 'tvec__min_df': 3, 'tvec__ngram_range': (1, 2), 'tvec__stop_words': 'english'}	0.983	0.970	0.966	0.972
2	Logistic Regression	CountVectorizer	{'cvec__max_df': 0.9, 'cvec__max_features': 2000, 'cvec__min_df': 3, 'cvec__ngram_range': (1, 1), 'cvec__stop_words': 'english'}	0.998	0.937	0.950	0.917
3	Logistic Regression	TfidfVectorizer	{'tvec__max_df': 0.9, 'tvec__max_features': 4000, 'tvec__min_df': 3, 'tvec__ngram_range': (1, 3), 'tvec__stop_words': 'english'}	0.988	0.964	0.965	0.958

Model Metrics (Test Data)

Cross-validation of training data: 0.980

Accuracy of test data: 0.974

Precision score: 0.978

Sensitivity score: 0.967

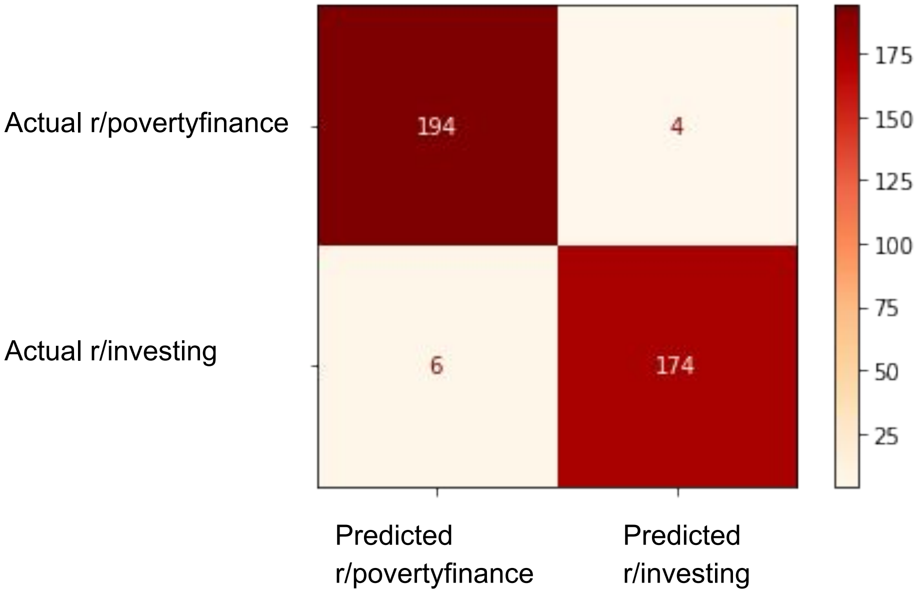
Model Metrics (Test Data)

Cross-validation of training data: 0.980

Accuracy of test data: 0.974

Precision score: 0.978

Sensitivity score: 0.967



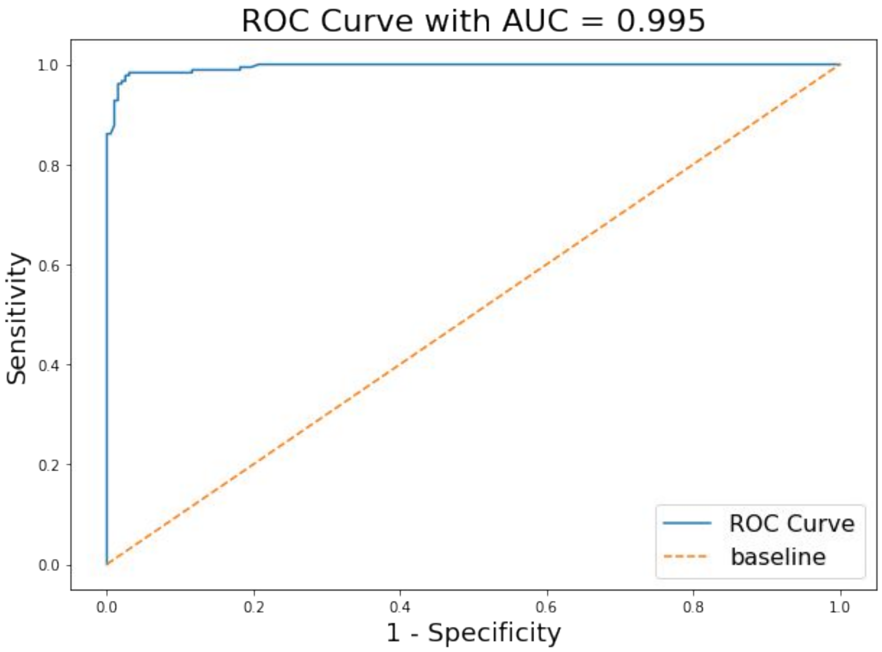
Model Metrics (Test Data)

Cross-validation of training data: 0.980

Accuracy of test data: 0.974

Precision score: 0.978

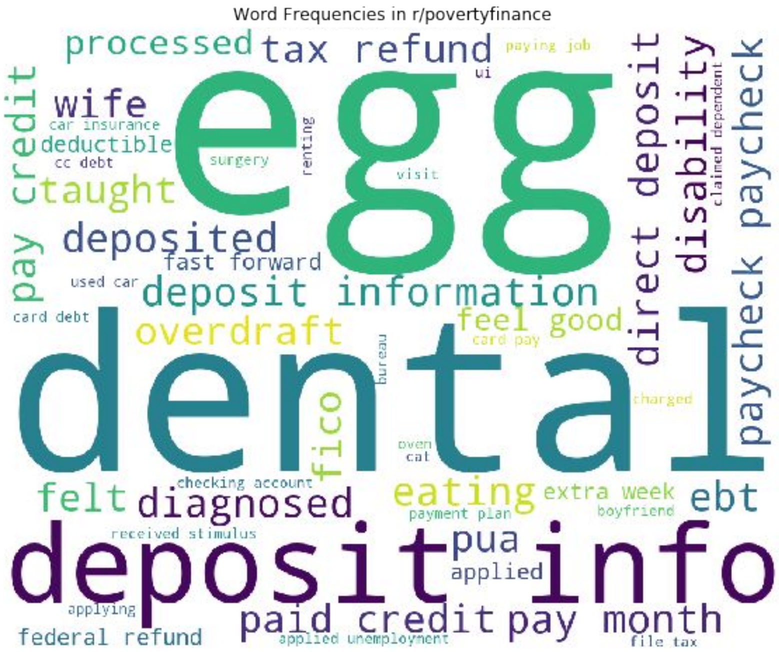
Sensitivity score: 0.967



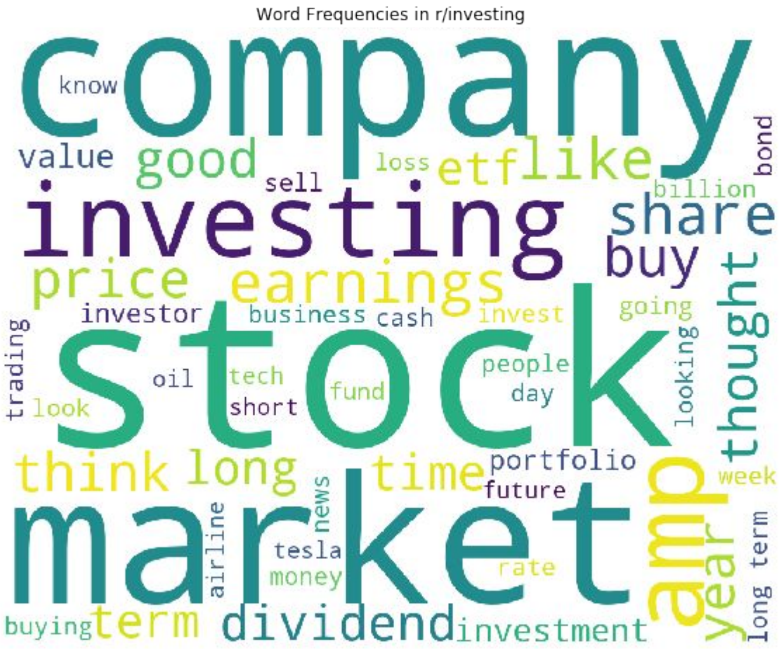
Conclusion & Recommendations

Key Findings:

r/povertyfinance


Word Frequencies in r/povertyfinance

r/investing


Word Frequencies in r/investing

Key Findings

- Individuals can identify subreddits to explore based on the below:
    - Topics in r/povertyfinance center around **saving up and restricting spending for the financially challenged**.
      money-saving tips / frugality, basic needs, insurance advice, paying debts, and living paycheck to paycheck.
    - Topics in r/investing center around **investing and making financial returns**.
      stocks, market outlooks, discussions on companies, and earning potentials.

Key Findings

- Financial Advisors / Personal Finance Blogs / Websites:
  - Identify what are the **hot topics** being discussed.

  - Such blogs and websites to **create more of such content**, especially times-sensitive ones in order to **improve engagement and visits**.

  - Financial advisors can **gain insights** on what their existing or potential clients would be interested in, to **improve their advisory services** to clients.

Key Findings

Misclassifications: words used were more generic across subreddits but different context

False Positives posts (predicted r/investing but actually r/povertyfinance)

free ***wendy*** *piece* ***nugget*** *friday*

index *housing cost specifically look lowest priced option index housing cost specifically look lowest priced option wondering example much would room cheap apartment cost generally data cost living look bare essential rather average spending*

False Negatives posts (predicted r/povertyfinance but actually r/investing)

anyone else notice seeking alpha ruined since week everything almost become ***paid*** *content feel like time delete app also month* ***ridiculous price***

Recommendations & Further Research

- **Removal of noise words** such as 'amp'. **Increasing n-grams** to get more context on more generic terms across subreddits.

- **Explore other tools: word2vec / stemming / other classifiers**

- **Increase our training dataset in size and of a longer timeline**

- **Explore relationships** between content, number of comments, and upvote ratios.

Thank you.