Reddit Web APIs & Classification

Roshan Khemlani

Agenda

- Problem Statement
- Background
- Data Collection
- Data Cleaning & EDA
- Preprocess & Modeling
- Evaluation
- Conclusion

Problem Statement

To classify Reddit posts from 2 different Subreddits, Personal Finance and Investing.

We will be using Natural Language Processing (NLP) and Classification modelling, to predict which subreddit a given post belongs to.

The model should provide researchers an understanding of how the average retail investors feel towards to allocating a certain portion of their income for investments.

They can work with financial consultants to identify which products/instruments are more marketable and strive to focus on that product during their sales pitch.

Background

The objective for the model is to find out how people from the middle class prioritize investing and what's holding them back.

With the information gathered from our analysis, we will notice;

- How they prioritize their income
- If there's any instrument that they are interested in
- If there's any obligations that's holding them back

Data Collection

Data was collected from the below subreddits API

https://www.reddit.com/r/investing/

https://www.reddit.com/r/personalfinance/

A total of 1742 rows was collected

- Investing - 842 rows
- Personal Finance - 897 rows

Data Cleaning & EDA

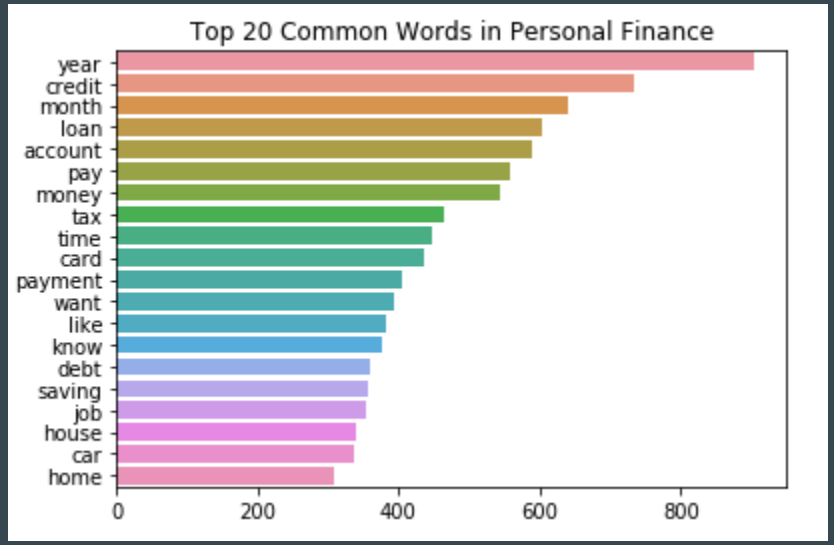
Data Cleaning

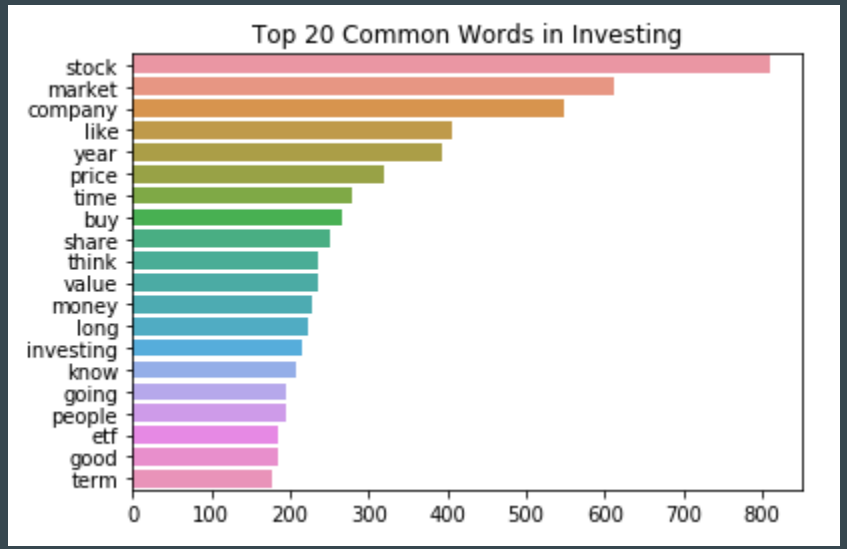
- Drop duplicate rows based on their selftext columns
- Filter null values and replace them with blank quotes (‘ ‘)
- Removed Urls, html tags, and everything that’s non-letters (numerics, line separators) from title and selftext columns.
- Excluded stopwords
- Created a new columns for the clean title and selftext, merged both investing and personal data sets.

Most Common Words in Investing

Most Common Words in Personal Finance

Top 20 Words Used in Personal Finance and Investing





Preprocess & Modeling

- Set up you train and test datasets

  X_train, X_test, y_train, y_test

- CountVectorizer

  Converting text data into a structured and numeric data frame

- TF-IDFVectorizer

  Transforms text to feature vectors that can be used as input to estimator.

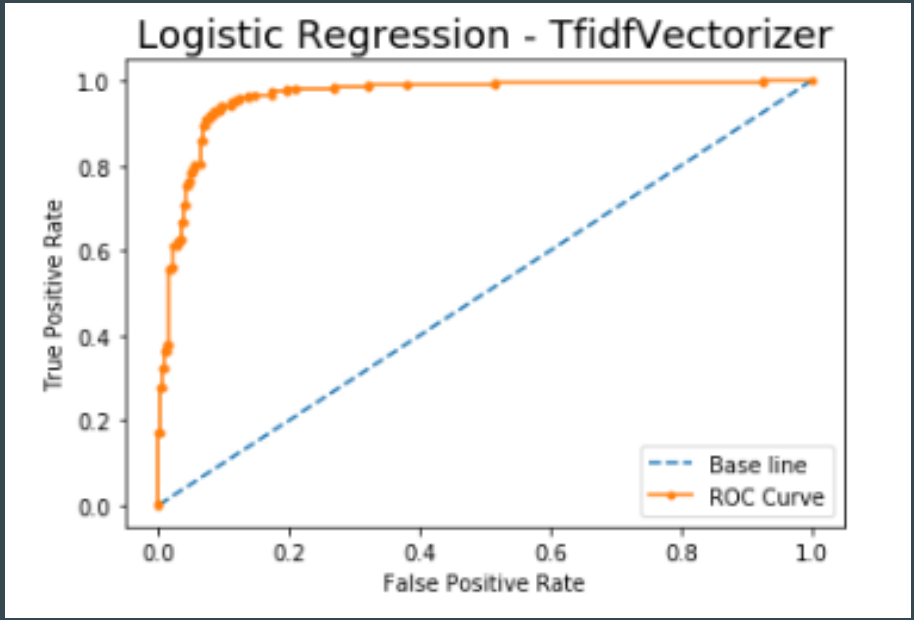
Modeling

Model	Train Score (CountVectorized)	Test Score (CountVectorized)	Train Score (TfidfVectorized)	Test Score (TfidfVectorizd)
Logistic Regression	0.9734	0.91134	0.9477	0.92
Naive Bayes (Multinomial)	0.9297	0.9165	0.9203	0.9113

Evaluation

Model	Accuracy	Sensitivity	Specificity	Precision
Logistic Regression (TfidfVectorized)	0.92	0.91039	0.92905	0.92363
Naive Bayes (CountVectorized)	0.91652	0.91039	0.92229	0.91696

ROC AUC Curve



ROC AUC Score: 0.963

Conclusion

The Naive Bayes with TfidfVectorizer and Logistic Regression with TfidfVectorizer worked very well with an high train and test score, even though both subreddits were technically related.

This could be attributed to the current market condition, with Covid-19 having a negative impact on job securities and financial stability.

Hence more retail investors, would not want to venture into riskier investments and focus to reduce their reliabilities.