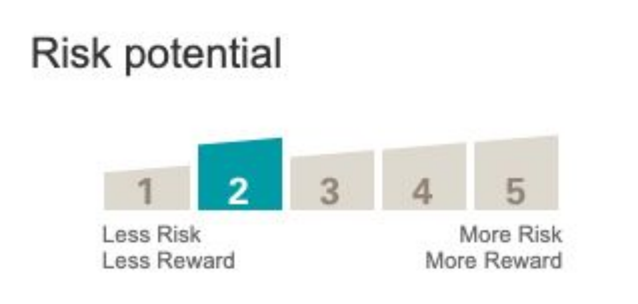
Investing For The People

r/ Bogleheads: Passive Investing For Lazy People

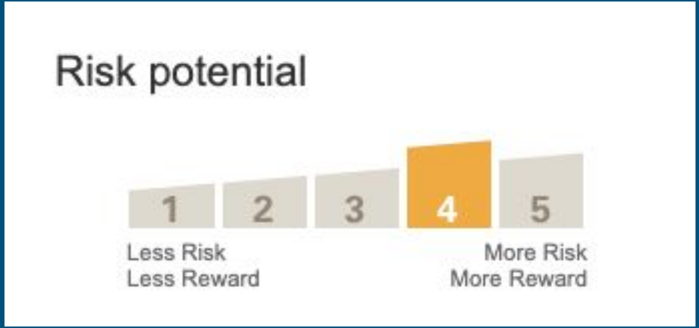
Using Natural Language Processing Can We Create a Classification Model That Predicts Whether A Post Belongs to r/Bogleheads Using Only The Features In The Title & Self Text Columns?

Subreddits

r/Bogleheads
Passive Indexing for Lazy Investors



r/Investing
Lose Money With Friends!



Data Collection & Preprocessing

Data Collection

-Using Pushshift's API We Collected Reddit Posts from r/Investing & r/Bogleheads were collected

-Approximately 1,000 submissions From Each Subreddit

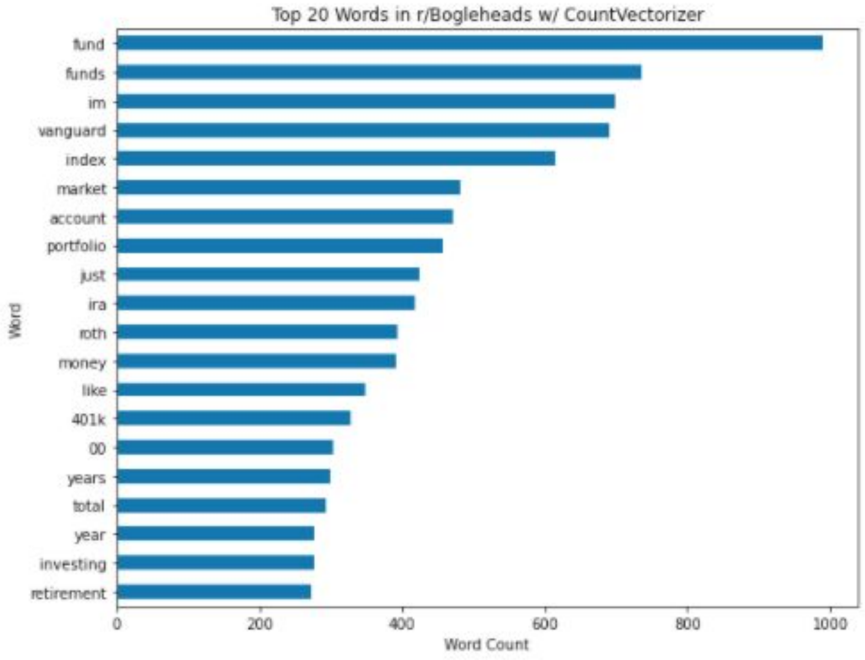
Data Cleaning/Preprocessing

- Remove Hyperlinks, Punctuation & Miscellaneous Characters

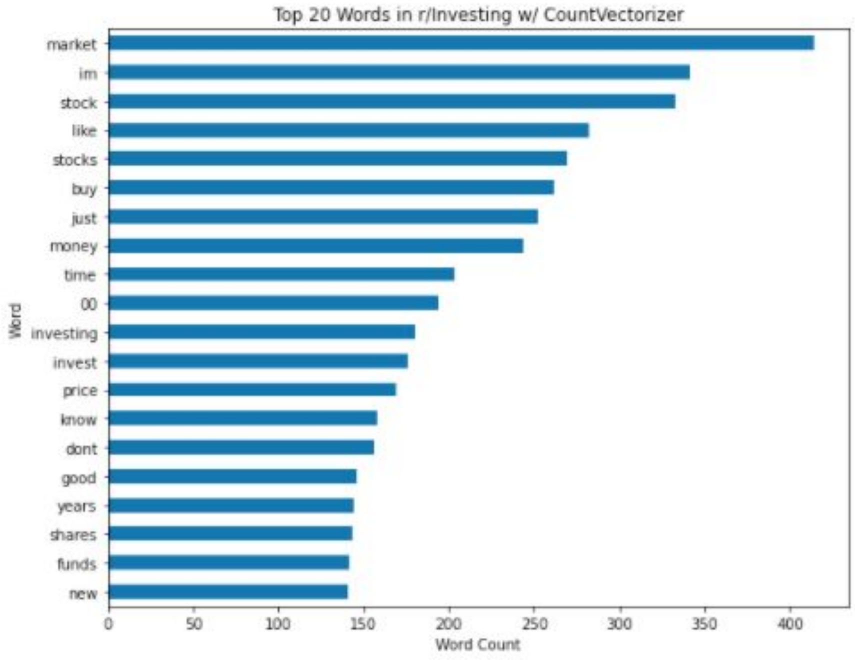
-Lemmatization Of Words To Convert To Base Words (Investing/Invested = Invest)

-Customized Stopwords List To Include Words That We Believed Would Not Help Classification (i.e. Removed)

Top Words By Subreddit



Top 20 Words in r/Bogleheads w/ CountVectorizer



Top 20 Words in r/Investing w/ CountVectorizer

Model Selection

- Logistic Regression + CountVectorizer
- Logistic Regression + TfidfVectorizer
- Gaussian Naive Bayes + CountVectorizer
- Gaussian Naive Bayes + TfidfVectorizer
- MultiNomial Naive Bayes + CountVectorizer
- MultiNomial Naive Bayes + TfidfVectorizer

Model Performance

Logistic Regression + TfidfVectorizer

-Training Score: 91.3%

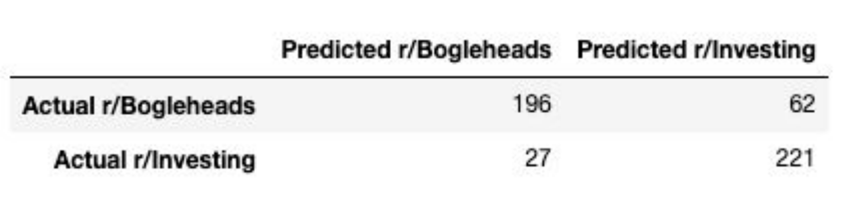
-Testing Score: 90.2%

-Max Features: 1000

- Only model where 1 & 2 gram tokens were used

Model Evaluation

Logistic Regression + TfidfVectorizer



	Predicted r/Bogleheads	Predicted r/Investing
Actual r/Bogleheads	196	62
Actual r/Investing	27	221

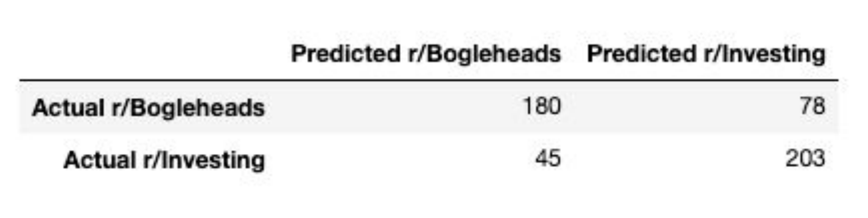
Accuracy: 82.4%

Misclassification: 17.6%

Sensitivity: 78.1%

Specificity: 76.3%

Gaussian Naive Bayes



	Predicted r/Bogleheads	Predicted r/Investing
Actual r/Bogleheads	180	78
Actual r/Investing	45	203

Accuracy: 75.7%

Misclassification: 24.3%

Sensitivity: 72.2%

Specificity: 69.7%

Misclassified Posts

Misclassified as r/Investing

"high yield corporate **bond** right now compared to equities it feels like **bond** are a better value bet has anyone looked into high yield corporate **bond** just wondering if it is worth a deeper look."

Misclassified as r/Bogleheads

“**Vti** **voo** **voog** i want to buy all is that a bad idea im young and all stocks”

Conclusion

- Logistic Regression + TfidFVectorizer Was The Best Performing Model
- Accuracy Score of 82.4% Achieved But Fell Short Of A Goal of 85%
- Identified Distinguishing Features For Each Subreddit
- Lots Of Overlap Of General Investing Terms Prevented Better Performance Of The Model

Top Coefficients for r/ investing

Word	Coef
Trading	1.917088
Company	1.822452
Stock	1.764724
Coronavirus	1.732456
Companies	1.622932
Oil	1.219998
Gold	1.162983
Price	1.44863
Robinhood	1.138840

Top 10 Coefficient for r/Bogleheads

Word	Coef
Vanguard	4.432238
Fund	3.254010
VTI	2.576931
VTSAX	2.518147
Funds	2.181807
Index	2.106754
IRA	2.057654
Bogleheads	1.949632
Portfolio	1.841194

*Adjusted for absolute value