DISTINGUISHING BETWEEN SPECULATIVE INVESTMENT TIPS

By James Salisbury

Problem Statement

- Differentiate between messaging focused on various types of investment recommendations
- Consider messages with a more speculative and risky connotation compared with safer and more grounded investments
- Optimize for sensitivity with the target being the more speculative messaging

Data Source - Reddit

r/WallStreetBets

The WallStreetBets subreddit is centered around positions for options and stocks. They have a no cryptocurrency, no penny stocks, no pump & dumps, no market manipulation, etc. posting guidelines. They are primarily focused on large bets of > $2,500 USD for options or > $10,000 USD for stocks.

r/Investing

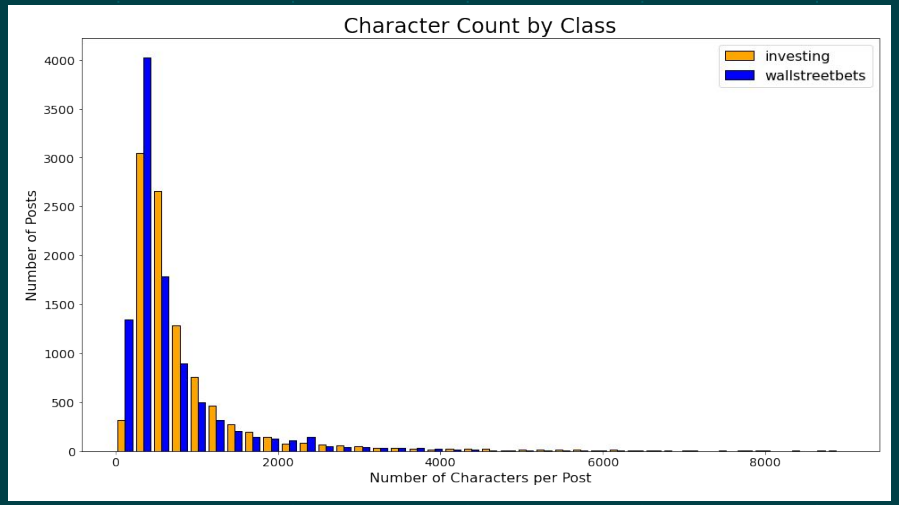
The Investing subreddit is focused on news items relevant to investors. It is meant to be an apolitical and all topics are meant to incite responses related to investing.

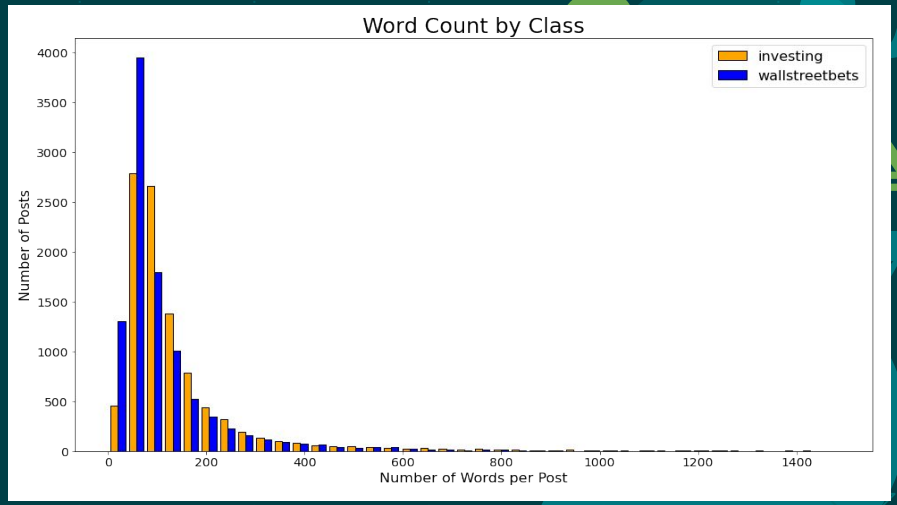
Data Description

- Minimum of 10,000 posts for each subreddit
- Collected the minimum posts while accounting for deleted and removed posts
- Pulled only self posts that included both a title and textual body for contents
- Deleted any duplicate posts, dictated by a post with the same author, title, and selftext

Data Discovery

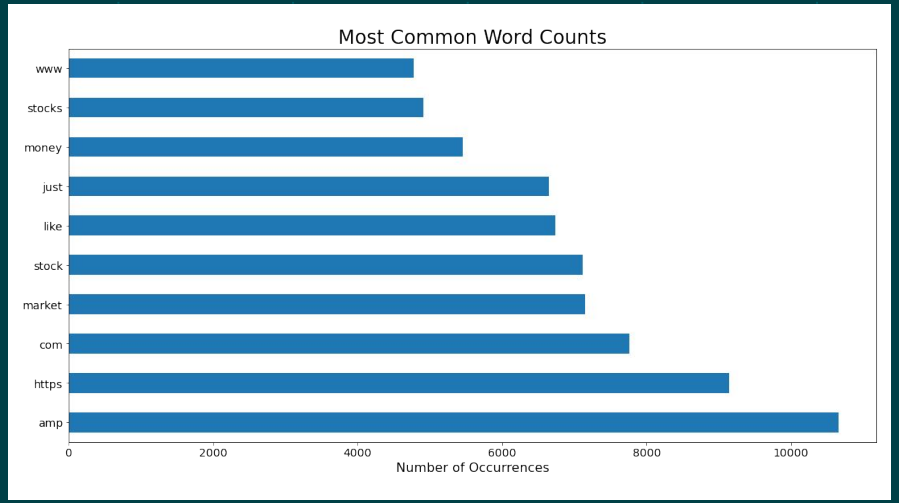
- WallStreetBets has a higher distribution of low character/word posts
- Investing has a higher distribution of higher character/word posts

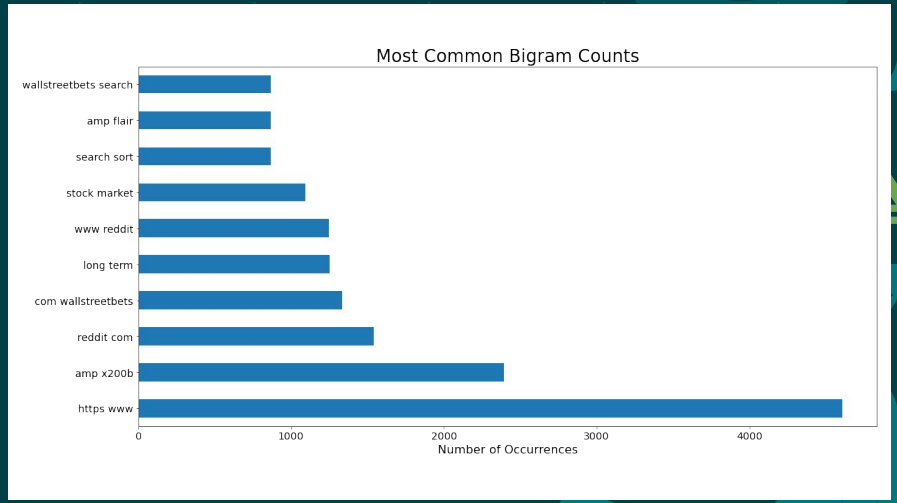




Data Discovery, cont

- After removing stopwords, evaluated commonly occuring words and phrases





Sentiment Analysis

Positive Investment Language

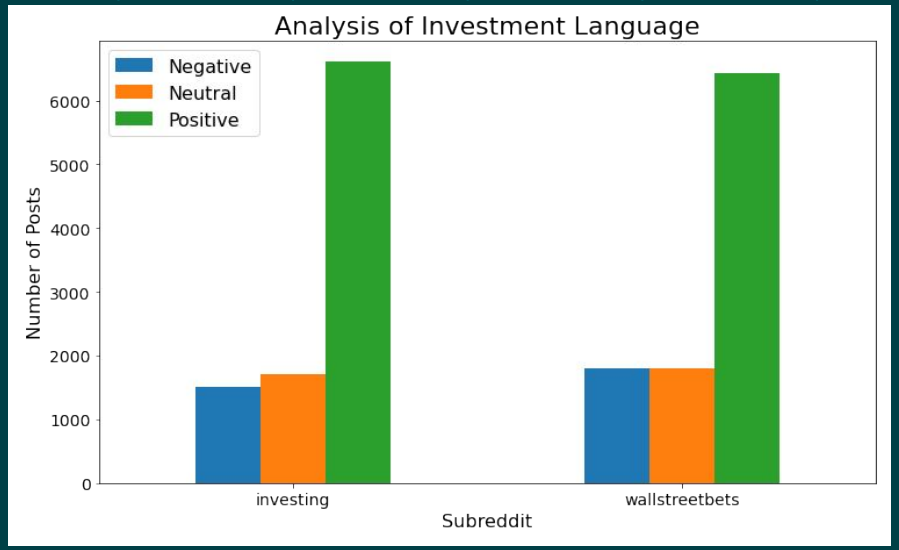
black, green, buy, bull, bullish, long, surplus, positive, up, gain, calls, profit, outperform, strong, best, opportunity

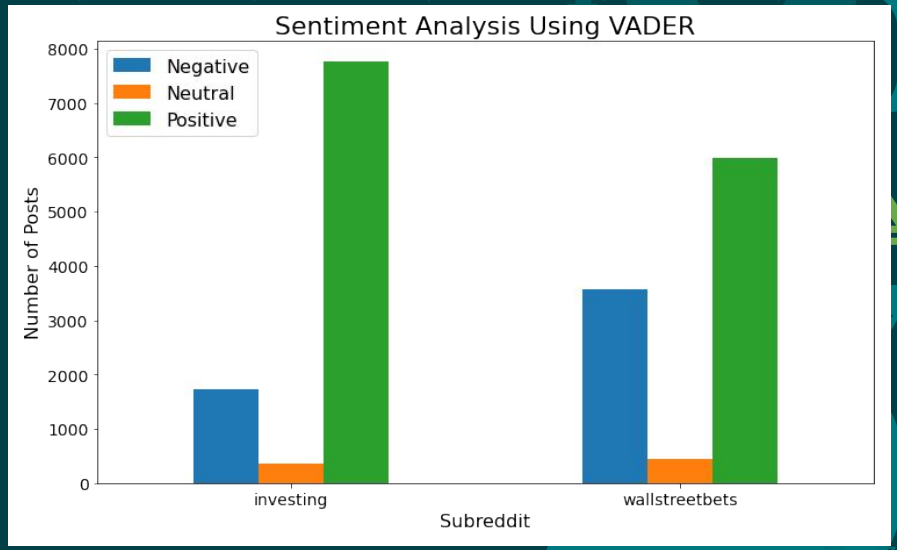
Negative Investment Language

red, sell, bear, bearish, short, debt, negative, decrease, drop, loss, puts, risk, uncertain, downside, overvalued, caution

Sentiment Analysis, cont

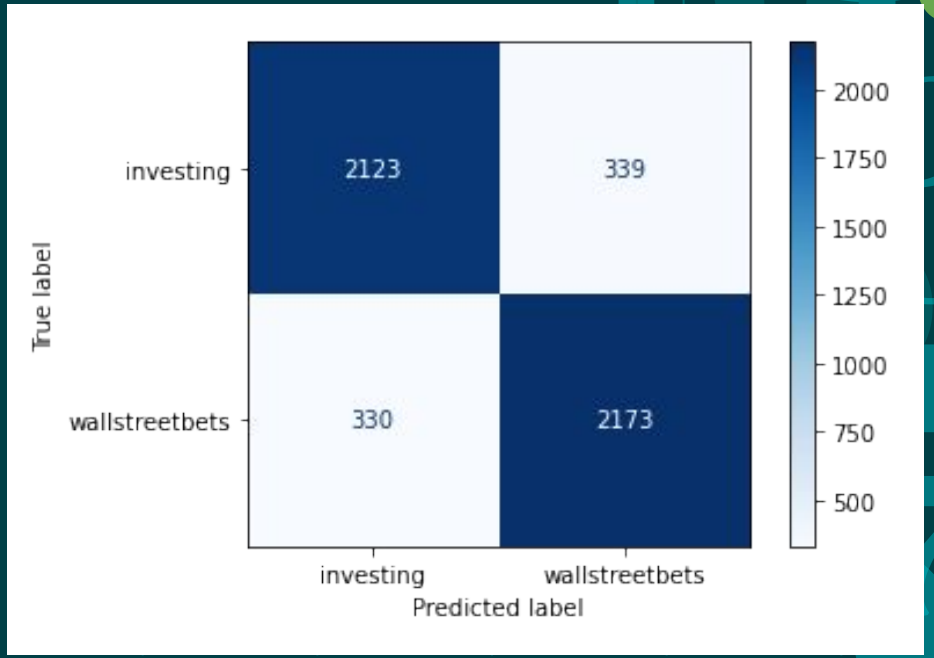
- Mainly positive sentiments for both subreddits and analyses
- Using the VADER dictionary, significantly lower neutral posts





Modeling

- Baseline of 50.4% for WallStreetBets
- Logistic Regression with text preprocessing
  - Vectorizing
  - Tokenizing
  - Stopwords
- Best score = 0.870
- Sensitivity score = 0.868



Conclusion

- Logistic Regression was the best model
  - Best sensitivity score
  - Discriminative
  - Collinearity
- Sufficiently classified WallStreetBets and Investing posts
- However, room for improvement

Next Steps

- Investigate any correlation between VADER and investment sentiment analyses
- Incorporate posts' comments to determine if the original post was looked upon favorably or not
- Further build out the investment term dictionary to allow for a deeper investment sentiment analysis

Thanks!

Any questions?