Subreddit Classification

With Web APIs and NLP

Ben Roberts - General Assembly

The burning question

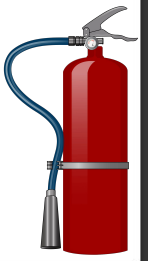
Given only the titles of posts from two subreddits, r/Stocks and r/Options, can we train a classifier on which subreddit a given post comes from using NLP and classification techniques?

r/Options

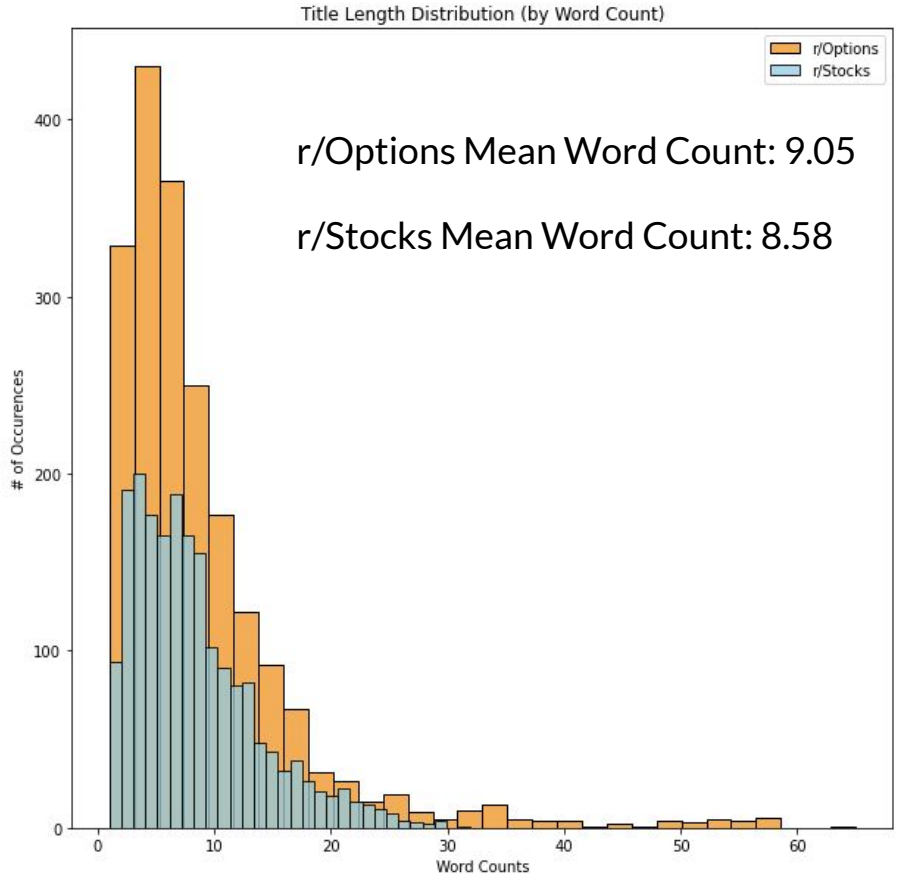
- **Created**
  October 2009
- **Members**
  837,000
- **Top Post**
  “The criminals that took GME down 371 points (77%) with only 8 million shares should rot in jail” (33.3k upvotes)

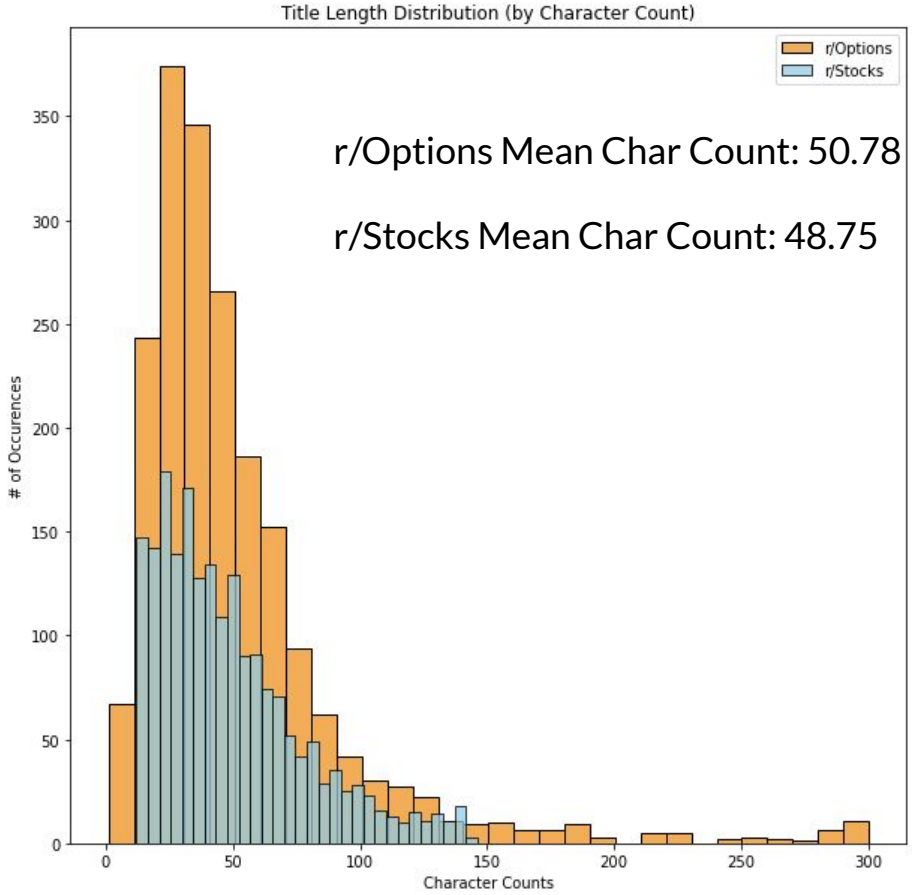
r/Stocks

- **Created**
  October 2009
- **Members**
  837,000
- **Top Post**
  “It's [redacted]ing awful seeing the "Silver" misinformation campaign everywhere I look” (102k upvotes)

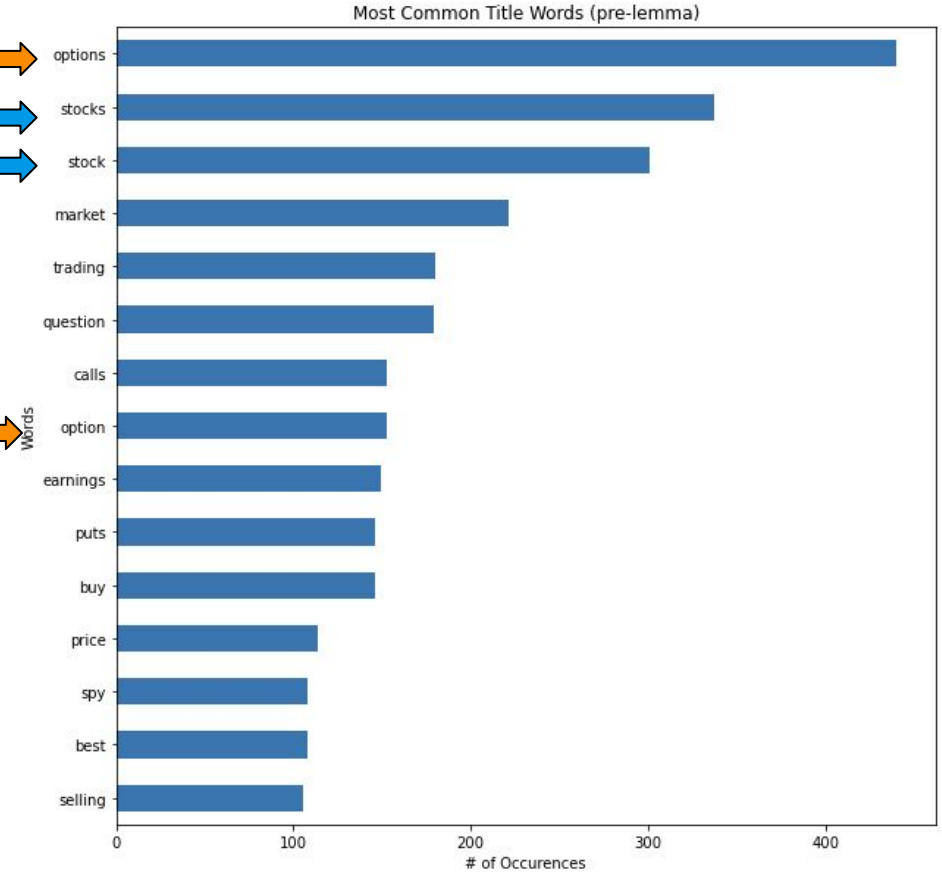


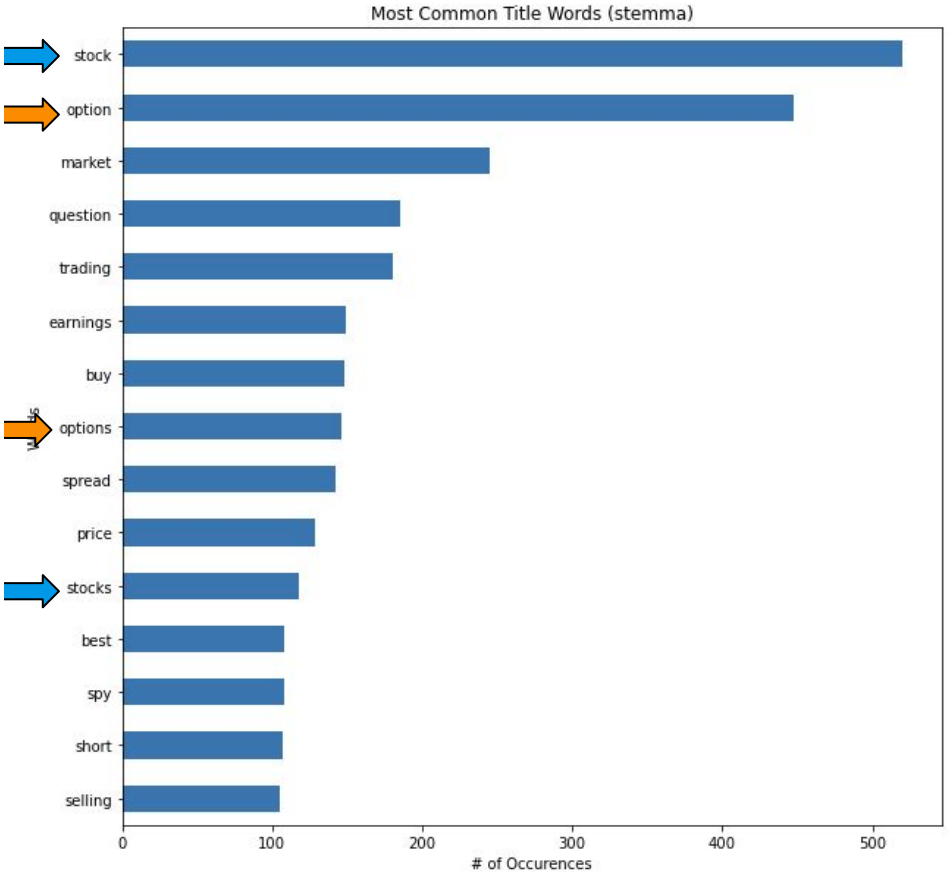
TITLE LENGTH by word count / character length



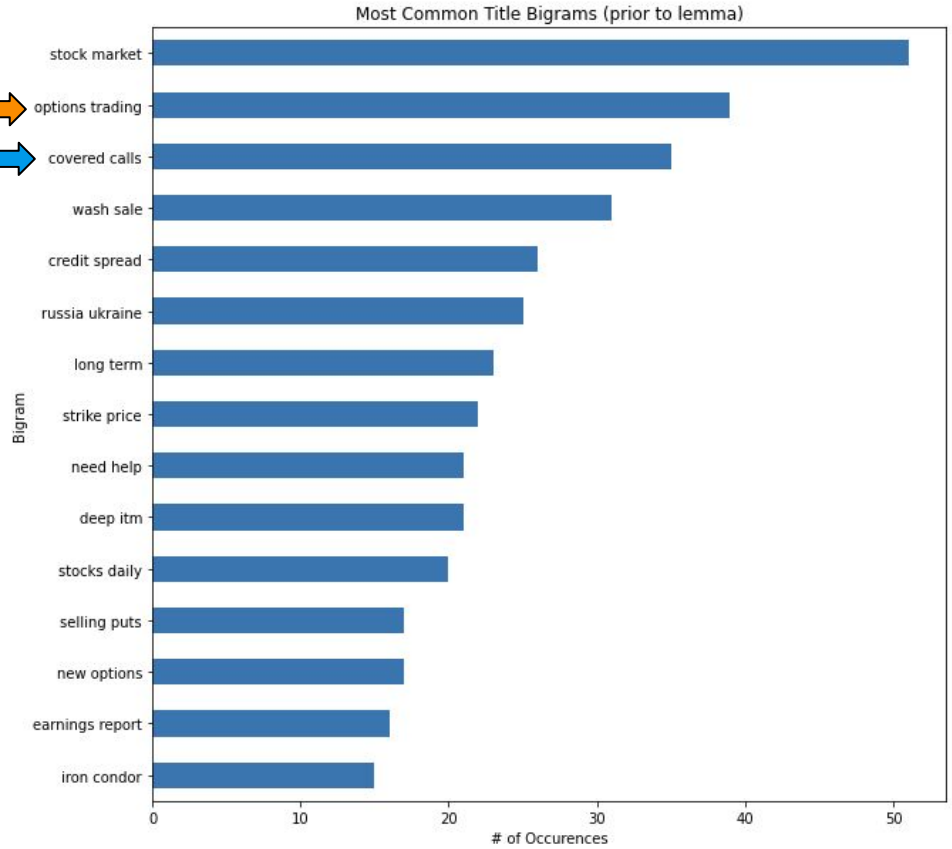


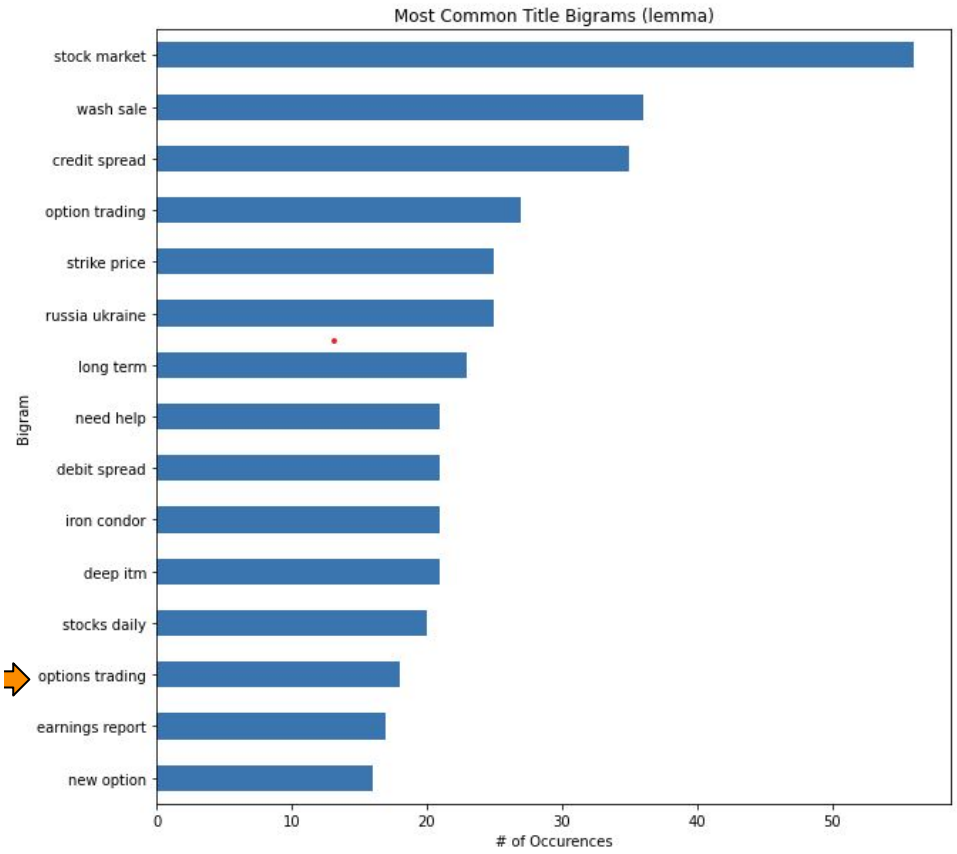
MOST COMMON WORDS combined



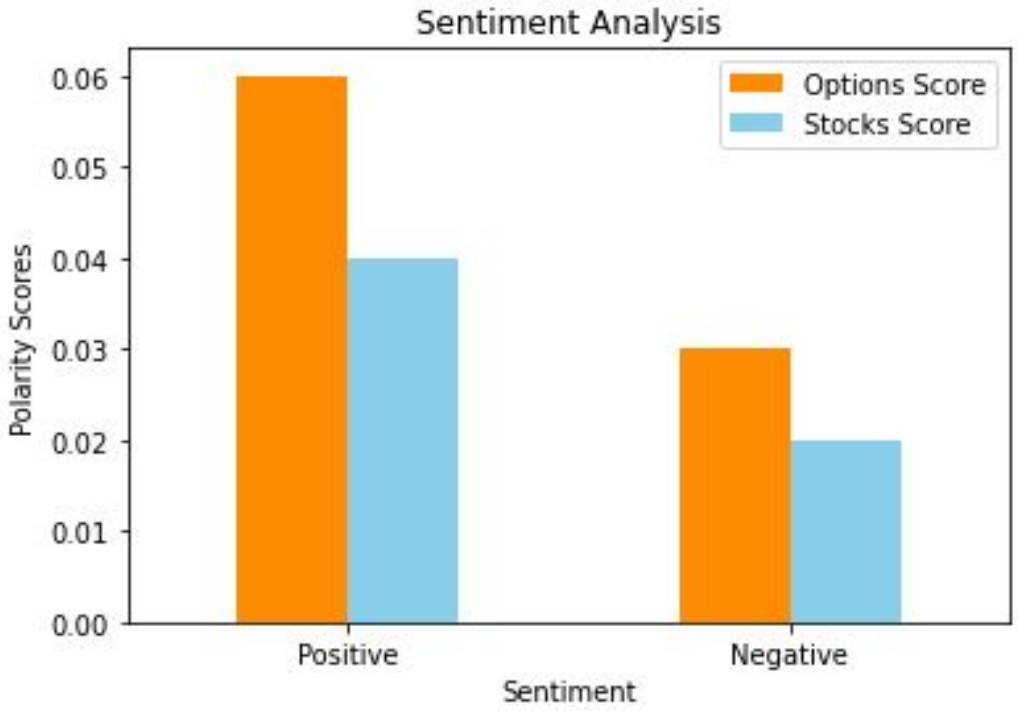


MOST COMMON BIGRAMS combined

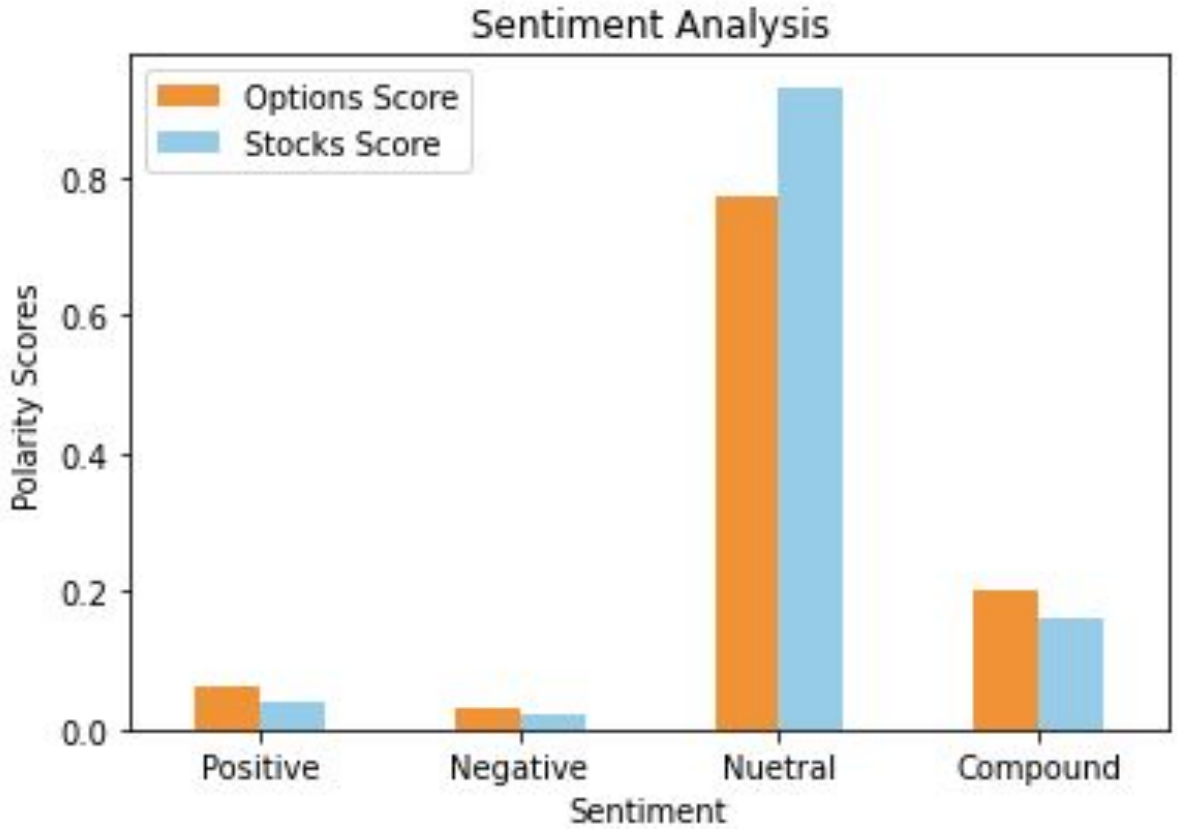




SENTIMENT ANALYSIS combined



SENTIMENT ANALYSIS combined



Baseline Score

Before any NLP or classification, our **BASELINE** model would predict that all posts originated from r/Options and the model would be correct 50.03%

Can we get any warmer?

These MODELS are hot!

Logistic Regression (82.8%)

Random Forest (83.7%)

K Nearest Neighbors (73.7%)

Best Parameters

LogReg: 'lr__max_iter': 1000, 'vect__ngram_range': (1, 2), 'vect__stop_words': None

Random Forest: 'vect__ngram_range': (1, 2), 'vect__stop_words': None

KNeighbors Best: 'vect__ngram_range': (1, 1), 'vect__stop_words': 'english'

Evaluation

Our model accurately predicted:

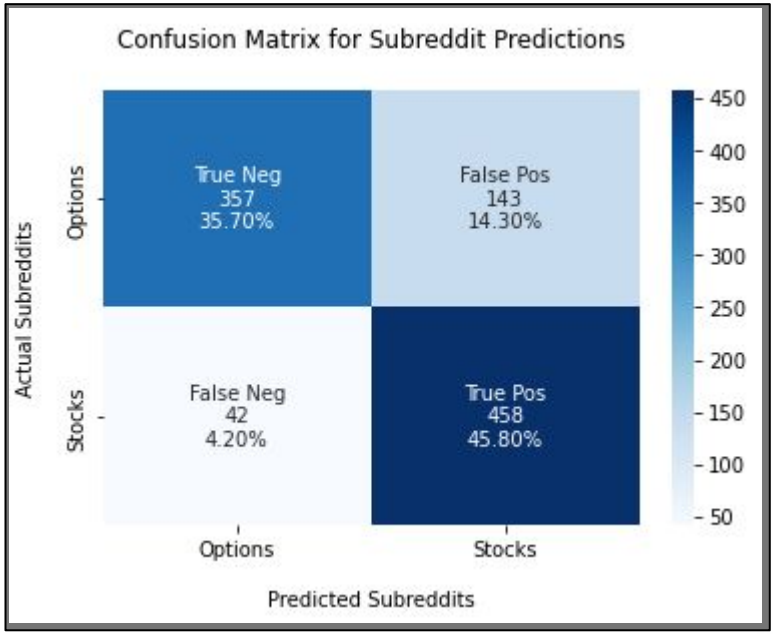
r/Options 357 times

r/Stocks 458 times

Our model inaccurately predicted:

r/Options 42 times

r/Stocks 143 times



Evaluation

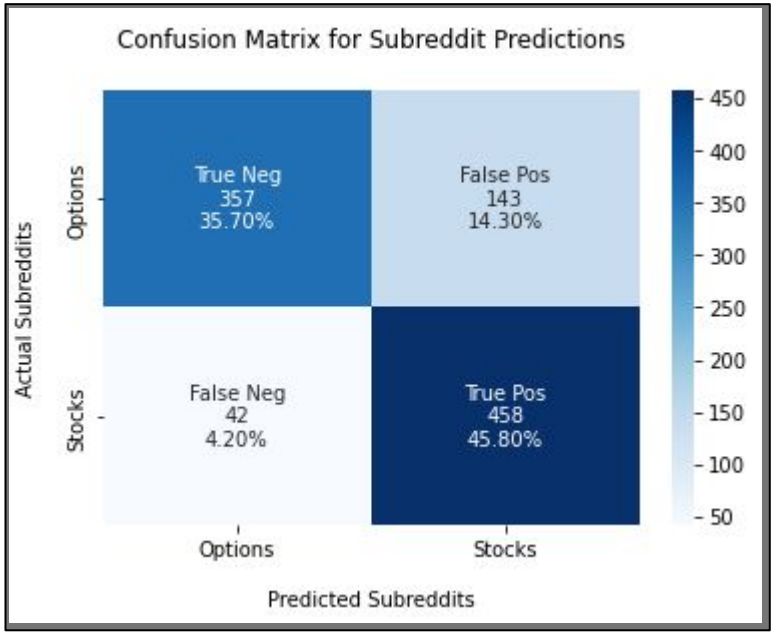
Accuracy: 81.5 %

Misclassification Rate: 18.5 %

Sensitivity: 91.6 %

Specificity: 71.4 %

Precision: 76.21 %



Recommendations

- **Date of posts**
  Hot topics and economic conditions change quickly
- **Comments**
  Hot stocks/option strategies and chances for arbitrage
- **Ticker Symbols and Sentiment**
  These may not be picked up using current Sentiment Analysis, lemmatizer, or stemmer

Conclusion

While there are a number of posts in foreign languages, many with no selftext, and a few different tactics we would like to employ in order to improve accuracy, the model is still accurate about 82% of the time in predicting the origin subreddit based on the lemmatized title of the post. This represents an improvement above the baseline of 32%.

Thank you!
Questions
and answers?