Project 3 NLP

Jiacheng Xu

Problem Statement

Natural Language Processing for subreddits

- 2 subreddits
  - CryptoCurrency
  - StockMarket

Crypto: 3.1 m members and 26.9 K active daily, created in 2013
Stock: 1.6 m members and 2.8 K active daily, created in 2008

Why do we select these two?

- Stock market and cryptocurrency market are both popular for quantitative finance workers.
- Cryptocurrecy is relatively new to investors, but many financial tools have already been widely used for Cryptocurrecy transactions.
- Online transaction platforms which used to serve for stocks, have also provide transaction service for cryptocurrency. Such as Robinhood.

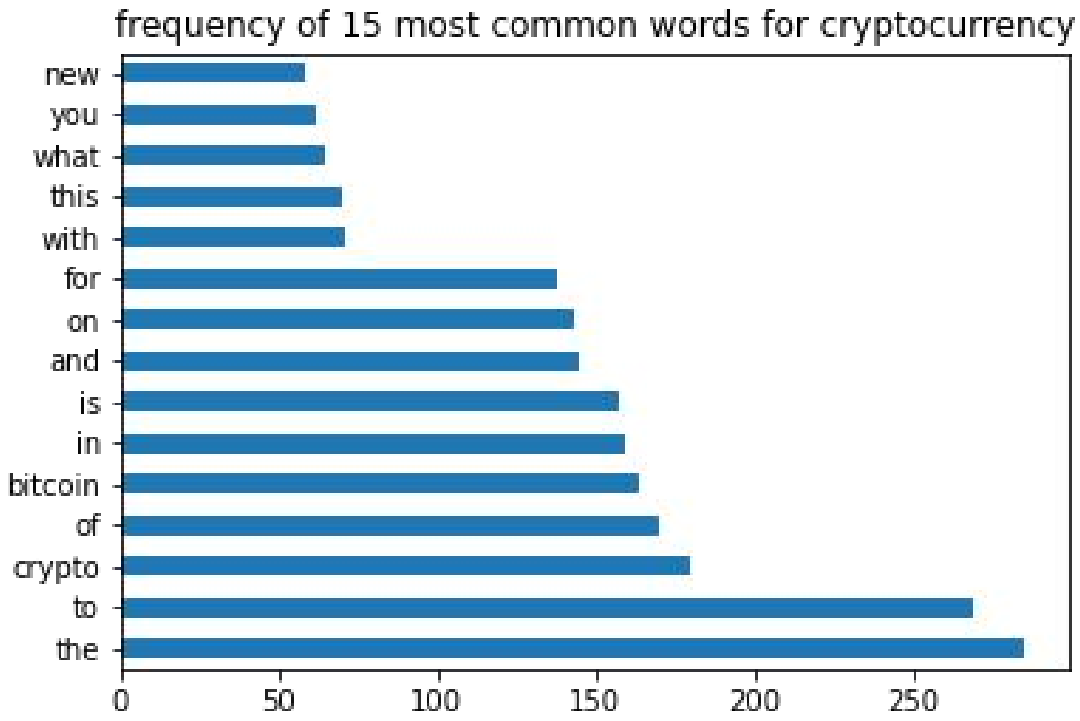
What do we do?

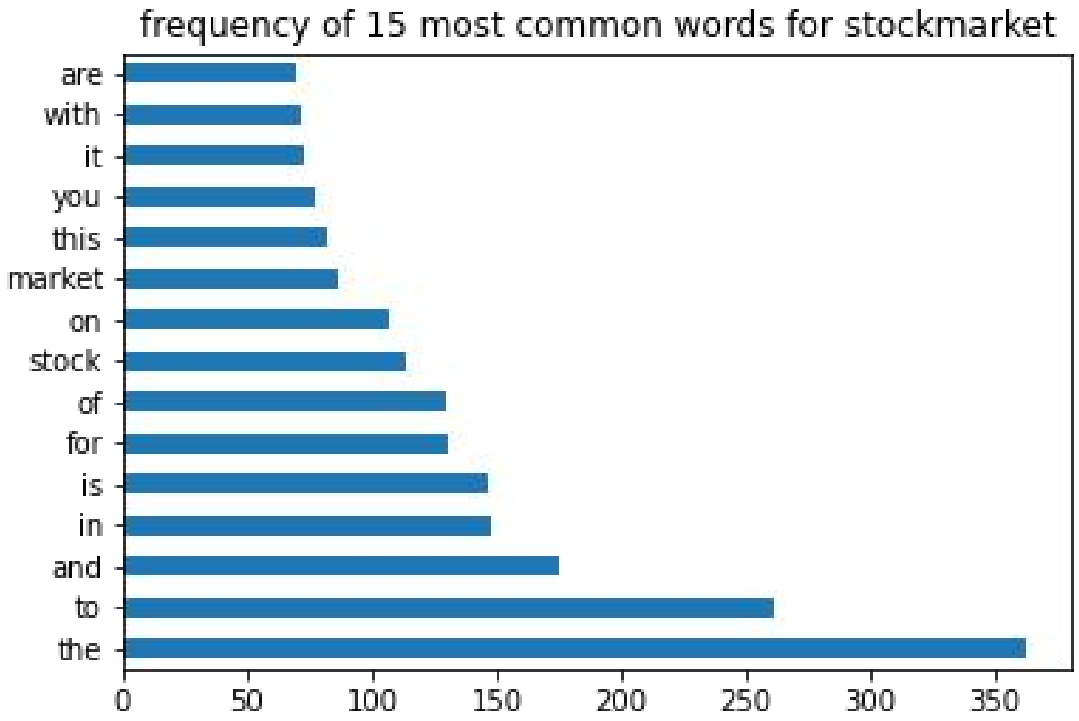
- Use NLP classifiers to predict which subreddit does the posts belong to.
- We want to make the predict more accurate, and better avoid severe overprediction.
- Use Pushshift's API to grab data from reddit.
  - 1000 posts for each subreddit
  - combine into a dataframe with 2000 posts' titles and the subreddit name which does them belong to.

Data collection

- Cryptocurrency data has 83 columns, stockmarket has 81 columns.
- Selftext column contains the description of a post. But most posts does not have a selftext.
- Score column contains the results of upvotes minus downsvote. Which implies how many viewers like one post.
- But we are only using the title column which contains the title of a post for classification

Exploratory Data Analysis





EDA

The authors who post most time in each subreddits:

'Mtraders': 26 posts in stockmarket

- like to use uppercase but does not appear to have special meaning.
- especially like to use 'wish'.

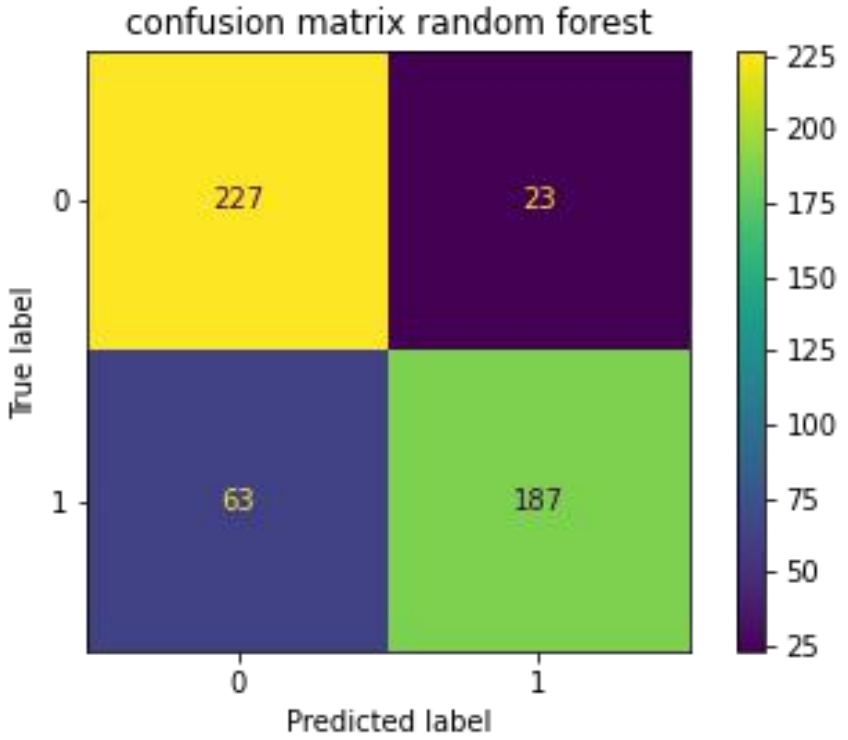
'nj_crypto_news': 22 posts in cryptocurrency

- no special pattern for word selection
- very clear pattern for word number

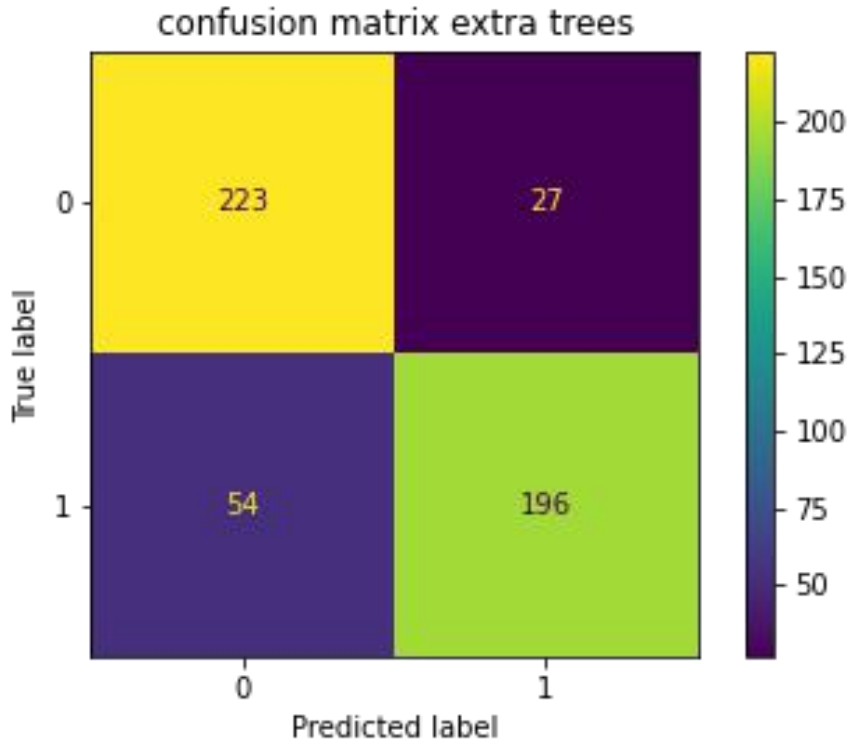
Model Selction

- Use Random Forest, Extra Trees, and SVM to classify NLP.
- CountVectorize my data.
- Parameter for random forest and extra trees
  - 'n_estimators': [100,150,200]
  - 'max_depth': [None, 1,2,3,4,5]
- Parameter for Kernel SVMs
  - 'C': np.linspace(1,5,20)
  - 'kernel': ['linear', 'poly', 'rbf', 'sigmoid']
  - 'degree': [2,3,4,5]

Random Forest and Extra Trees

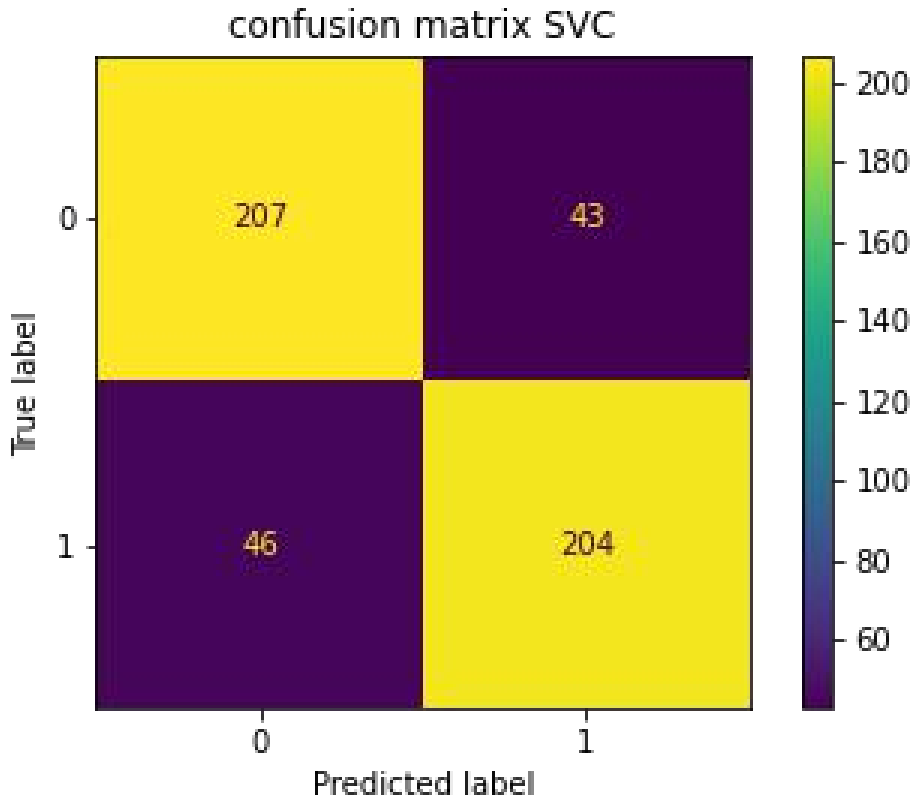


(0.998666666666667, 0.828)



(0.998666666666667, 0.838)

Kernel SVMs



(0.976, 0.822)

Conclusion

- All my classification models are overpredicting.
- Random Forest and Extra Trees are predicting the most accurately with high variance
- SVMs is predicting slightly worse but has a lower variance
- Random Forest and Extra Trees both have high precision score for predict Cryptocurrency
- SVMs doing well for predicting both subreddits.