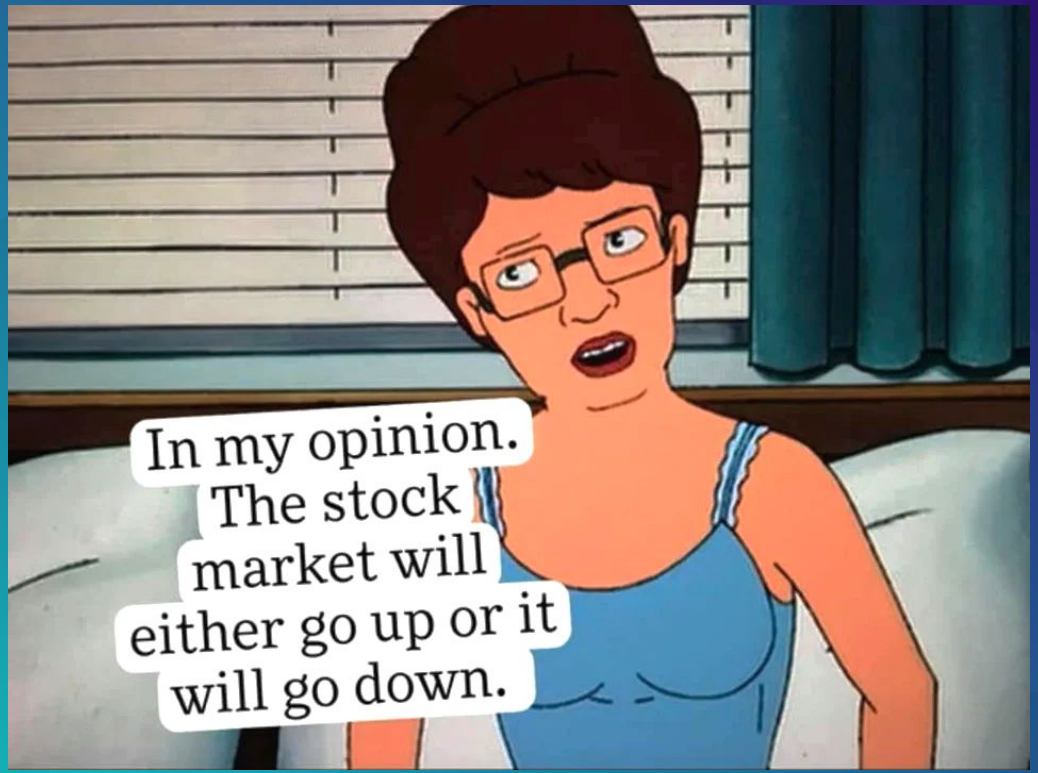
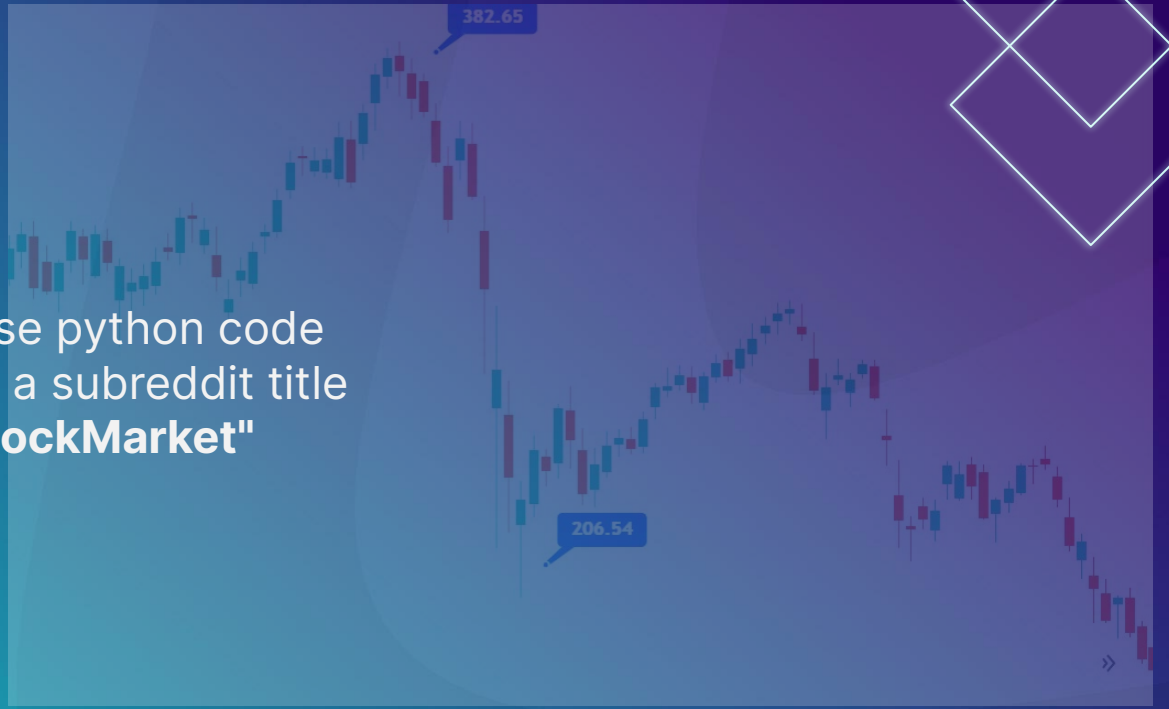
NLP CLASSIFICATION SUBREDDITS

Julian Chang

‘STOCKS’ VS ‘STOCKMARKET’

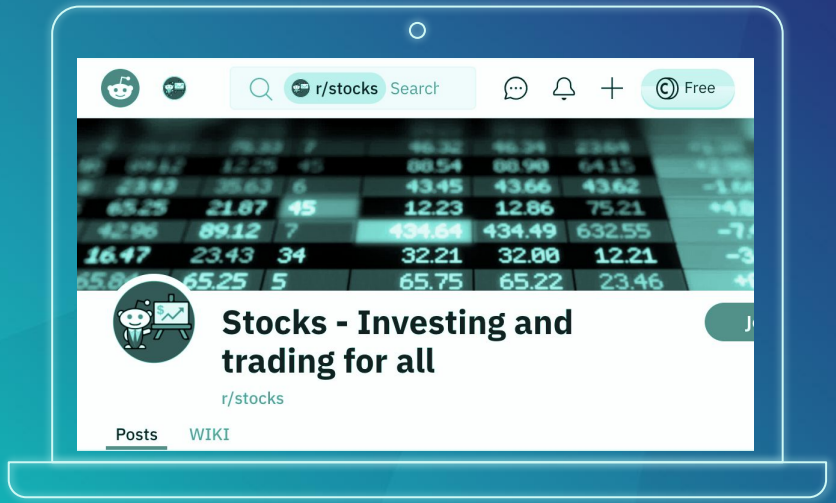


PROBLEM STATEMENT



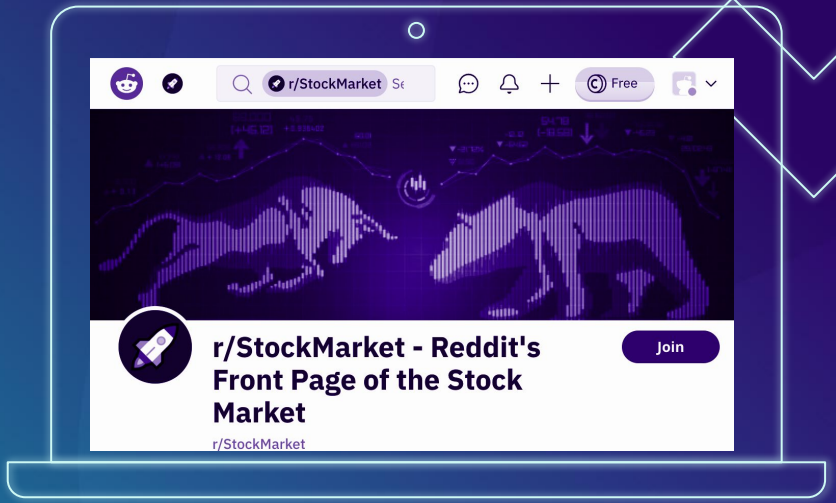
Goal of this project is to use python code **classifiers** to determine if a subreddit title belongs to **"stocks"** or **"StockMarket"**

SAMPLE OF PHRASES



stocks

'Will Beijing Supersede Hong Kong?'



StockMarket

'How to Value a Company with Multiples'

15130 *Data points extracted with*
Pushshift API

STOCKS

STOCKMARKET

TRAINING MODEL

PARAMETERS

DATA

15130 data points
Evenly from both subreddits

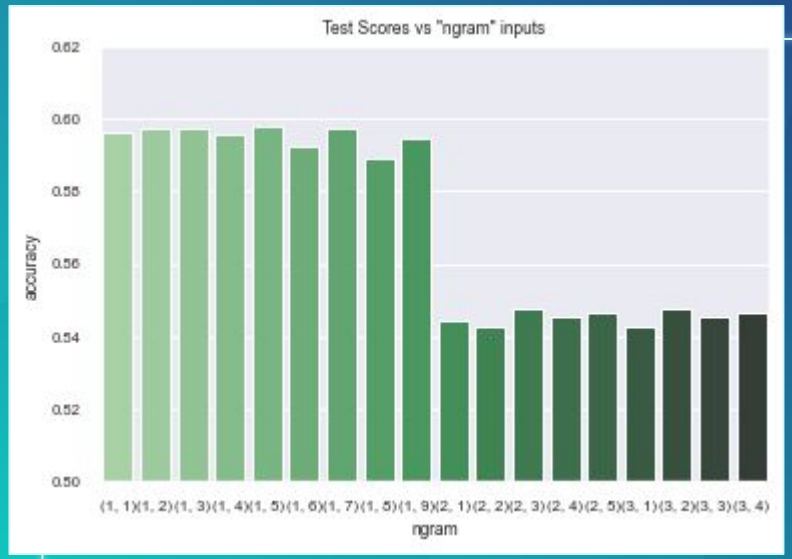
TOKENS

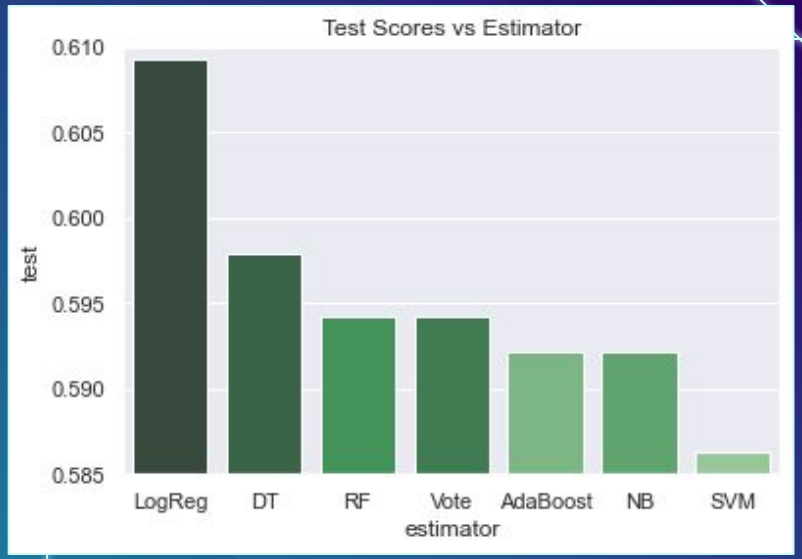
2
CountVectorizer,
TfidfVectorizer

ESTIMATOR

6 models inc.
NB, RF, Adaboost, VotingClassifier, SVM, LogReg

NGRAM AND ESTIMATORS





COUNTVECTORIZER()

(1,5)

NGRAM

Based on trials

3K

FEATURES

Logistic Regression can handle many features

ENGLISH

STOP WORDS

Pre-determined list of stop words is used

RANDOMFORESTCLASSIFIER()

5

MAX DEPTH

Low tree depth to reduce overfitting

200

ESTIMATORS

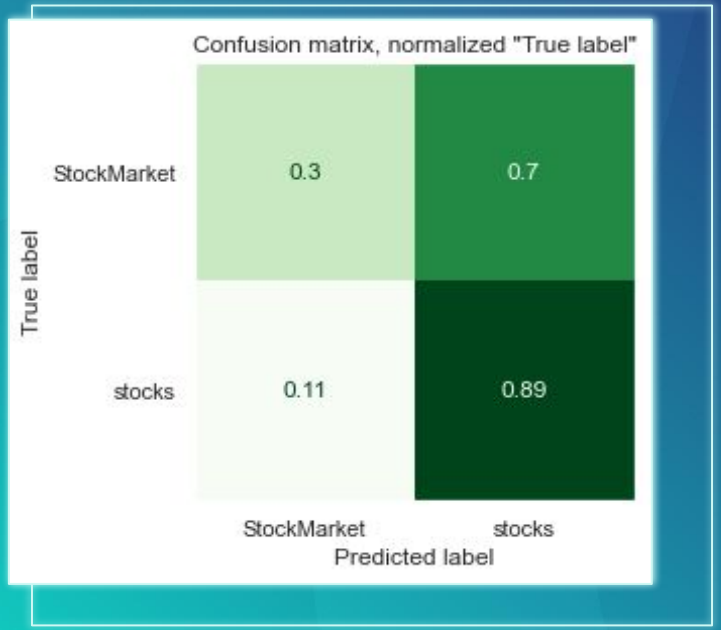
Reduce the compute time on the large dataset

random seed: 42, n_jobs: -1

GRADING MODEL

RANDOM FOREST

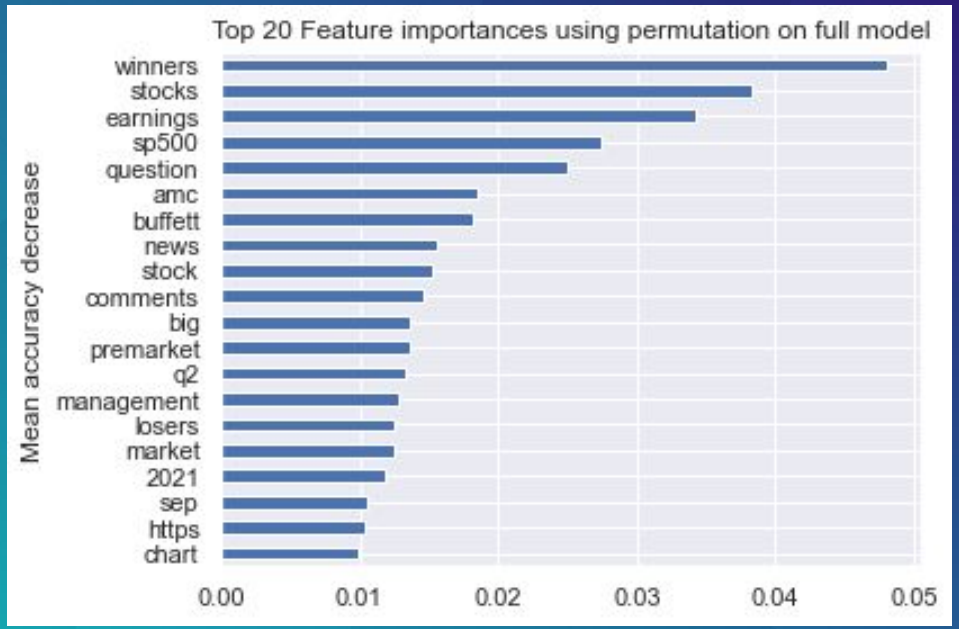
“Better than a coin flip”



SCORES

Accuracy	60%
Sensitivity	88%
Specificity	30%

RANDOM FOREST - FEAT IMPORTANCE



CONCLUSION

MODEL

Random Forest is the best trialled model

ACCURACY

Model can be useful to up to 60% of the time

RANKING

Feature Importance is prescriptive

SMART

Is the model better than human classification?

THANKS!

Do you have any questions?

changjulian17@gmail.com
github.com/changjulian17

/julian-chang/