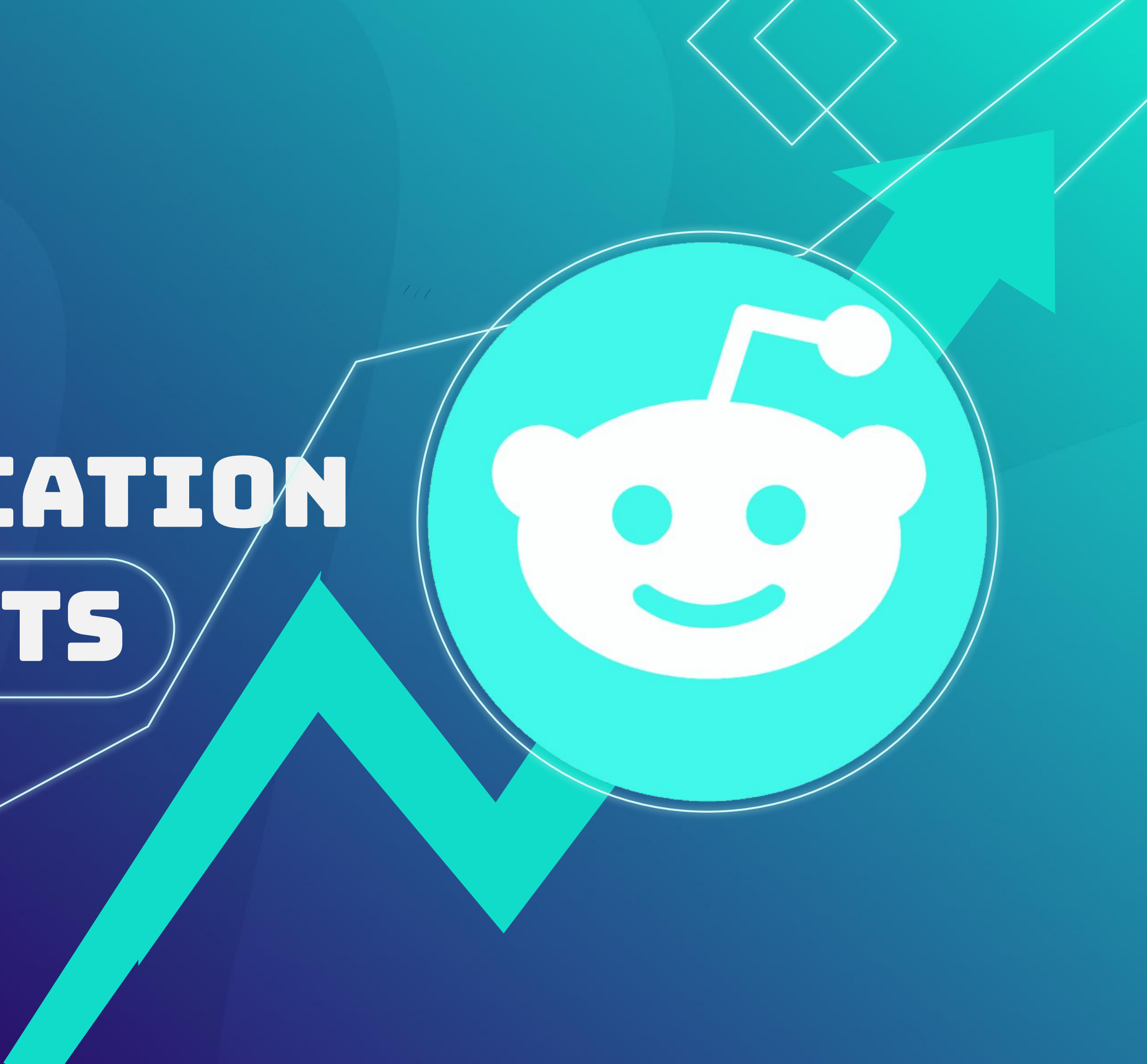
# NLP CLASSIFICATION SUBREDDITS

Julian Chang



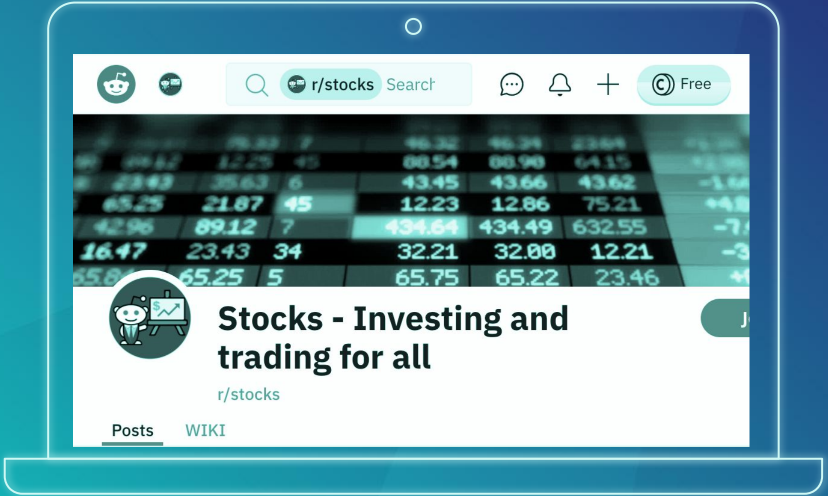
# ‘STOCKS’ VS ‘STOCKMARKET’

In my opinion. The stock market will either go up or it will go down.

# PROBLEM STATEMENT

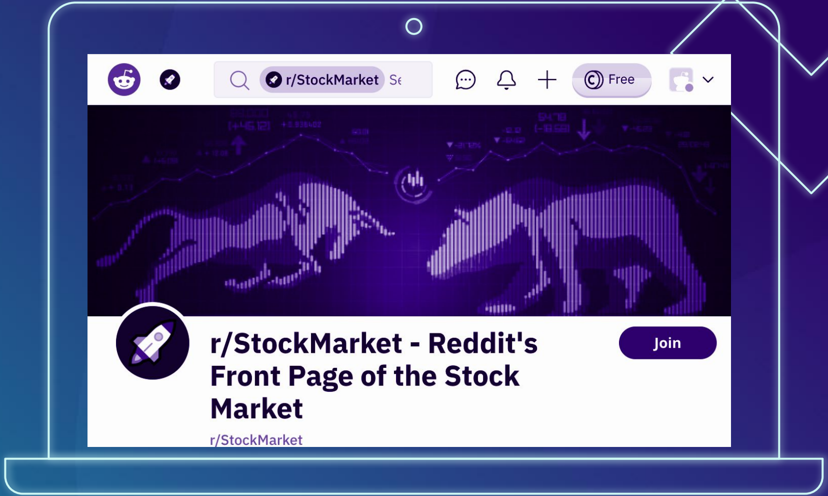
The Goal of this project is to use python code classifiers to determine if a subreddit title belongs to "stocks" or "StockMarket"

# SAMPLE OF PHRASES



Stocks - Investing and trading for all

r/stocks



r/StockMarket - Reddit's Front Page of the Stock Market

r/StockMarket

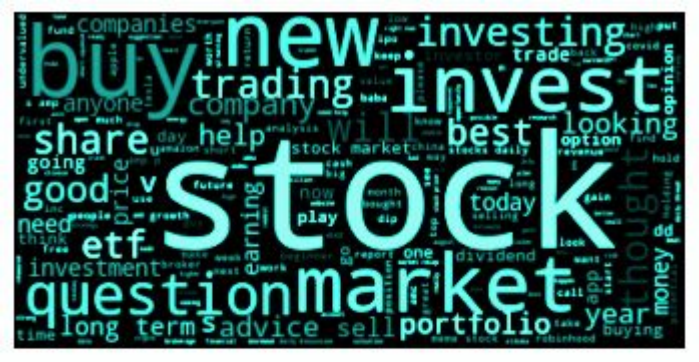
stocks

'Will Beijing Supersede Hong Kong?'

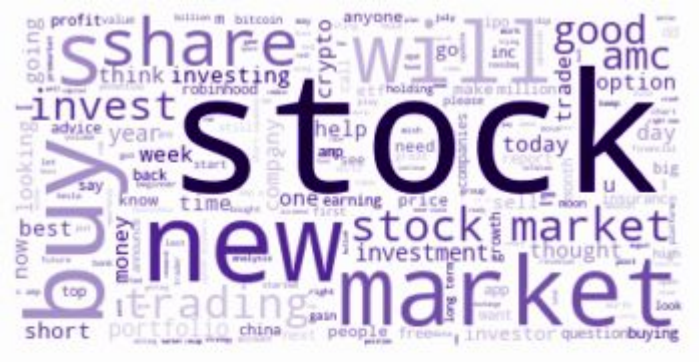
StockMarket

'How to Value a Company with Multiples'

## STOCKS



## STOCKMARKET



# TRAINING

# MODEL

# PARAMETERS

## DATA

15130 data points
Evenly from both subreddits

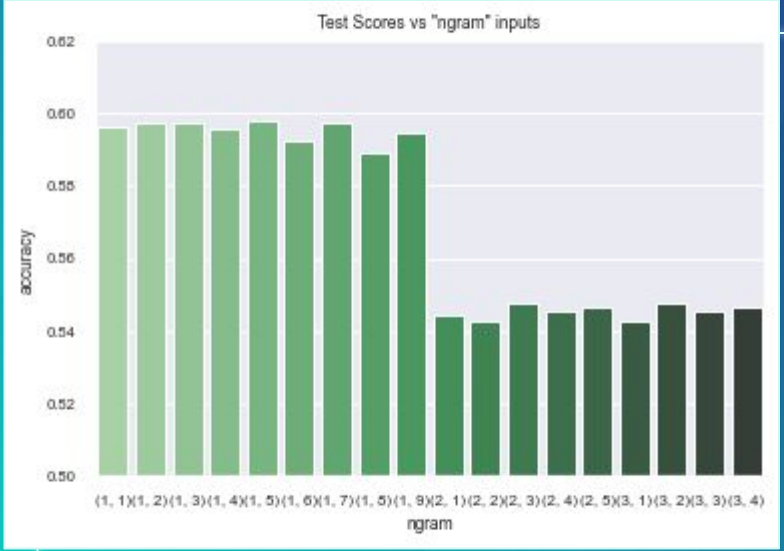
## TOKENS

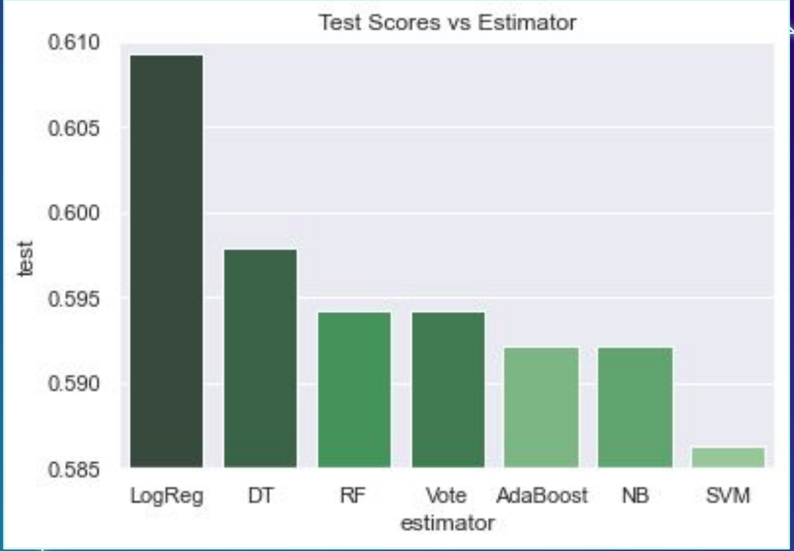
CountVectorizer,
TfidfVectorizer

## ESTIMATOR

NB, RF Adaboost,
VotingClassifier, SVM,
LogReg

# NGRAM AND ESTIMATORS

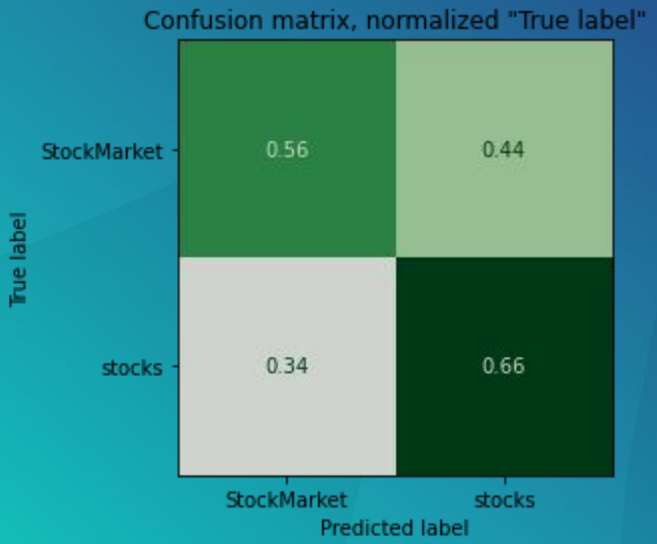




# GRADING

# MODEL

# LOGISTIC REGRESSION


Confusion matrix, normalized "True label"

## SCORES

Accuracy 61%

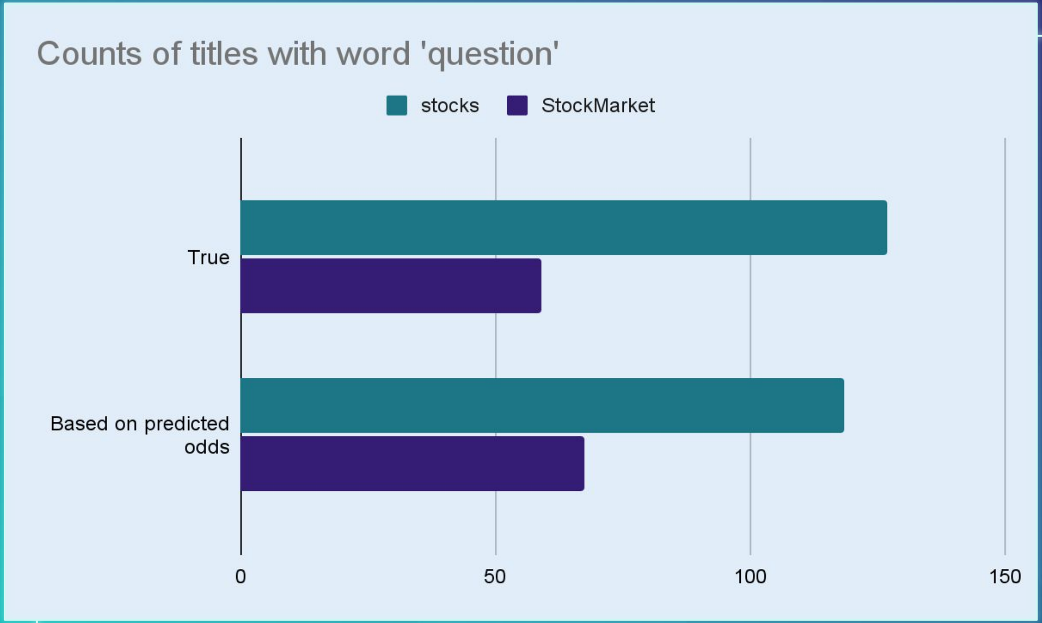
Sensitivity 66%

Specificity 56%

# LOGISTIC REGRESSION - COEFFICIENTS

| | TOP 3 COEFFICIENTS | |
|---|---|---|
| FEATURE | COEFFICIENT | ODDS |
| STOCKS | 0.57 | 1.76 |
| QUESTION | 0.56 | **1.75** |
| ADVICE | 0.53 | 1.70 |

# LOGISTIC REGRESSION - COEFFICIENTS



Based on data there is 2.15 times as many `stocks` titles with the word “question”

Every word "question" in a subreddit title is 1.75 times as likely to be considered from `stocks` subreddit

# CONCLUSION

## MODEL

Logistic Regression is the best trialled model

## ACCURACY

Model can be useful to up to 61%

## RANKING

model coefficients understandable

## SMART

Is the model better than human classification?

# THANKS!

Do you have any questions?

changjulian17@gmail.com
github.com/changjulian17

/julian-chang/

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

**Please keep this slide for attribution**