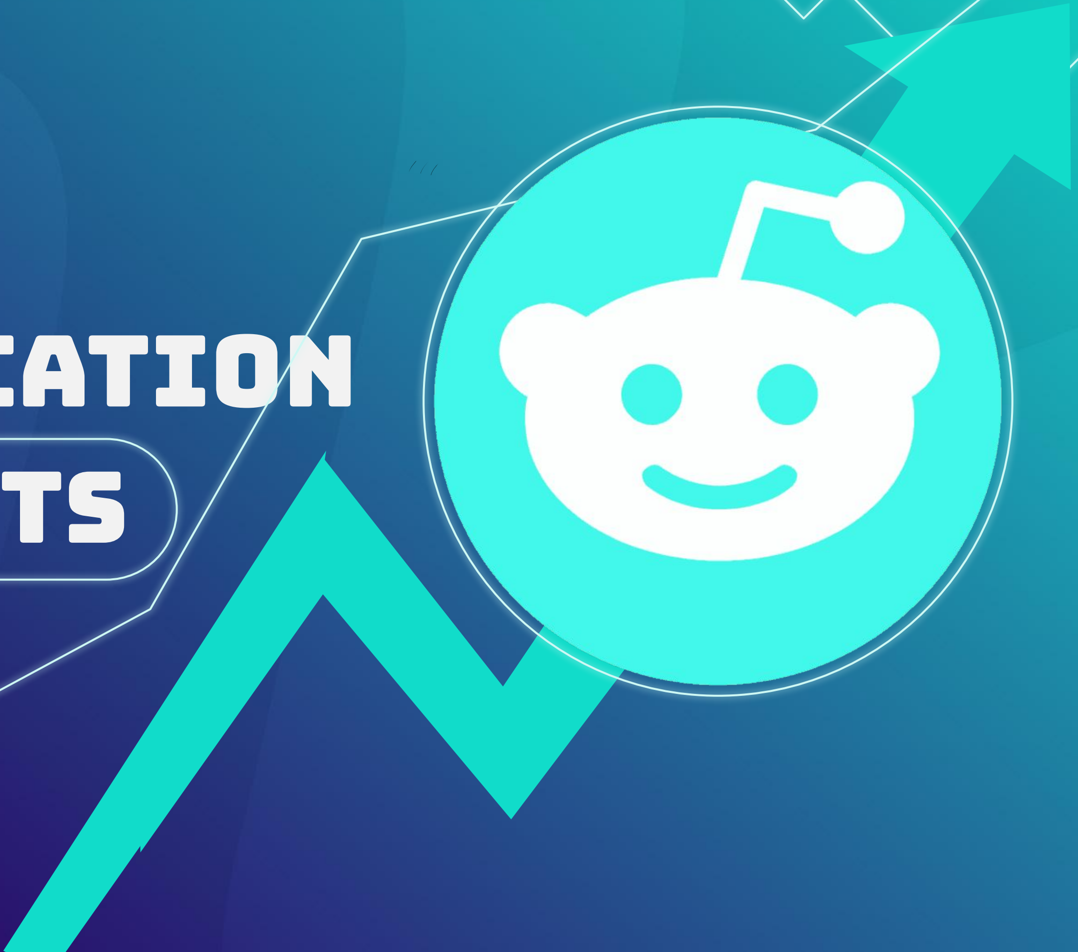
# NLP CLASSIFICATION SUBREDDITS

Julian Chang

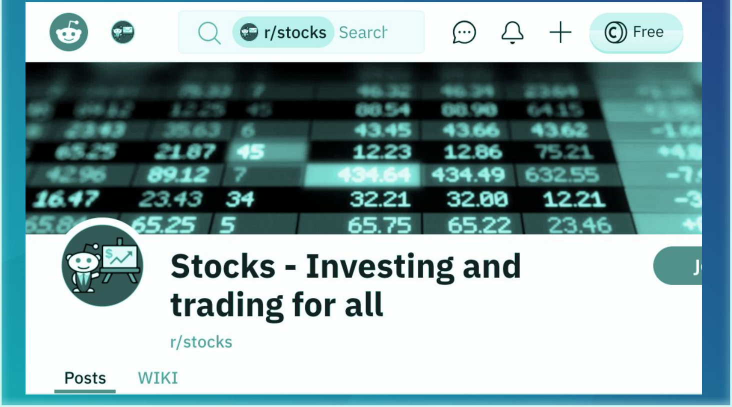


# ‘STOCKS’ VS ‘STOCKMARKET’

# PROBLEM STATEMENT

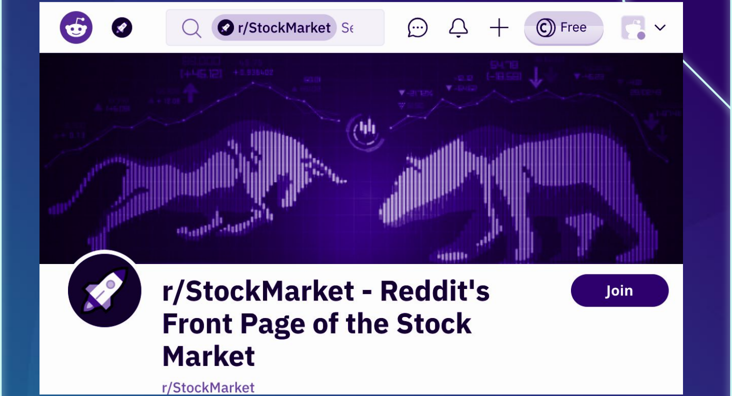
Goal of this project is to use python code **classifiers** to determine if a subreddit title belongs to **"stocks"** or **"StockMarket"**

# SAMPLE OF PHRASES



Stocks - Investing and trading for all

r/stocks



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

r/StockMarket

stocks

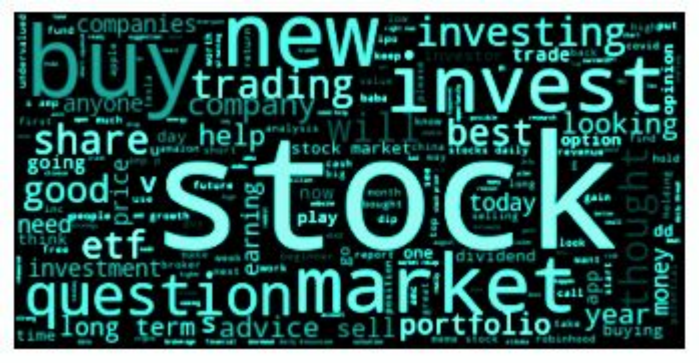
'Will Beijing Supersede Hong Kong?'

StockMarket

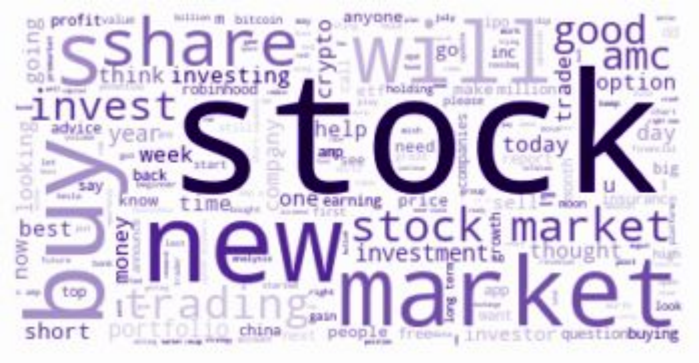
'How to Value a Company with Multiples'

***15130** Data points extracted with Pushshift API*

## STOCKS



## STOCKMARKET



# TRAINING

# MODEL

# PARAMETERS

## DATA

**15130** data points
Evenly from both subreddits

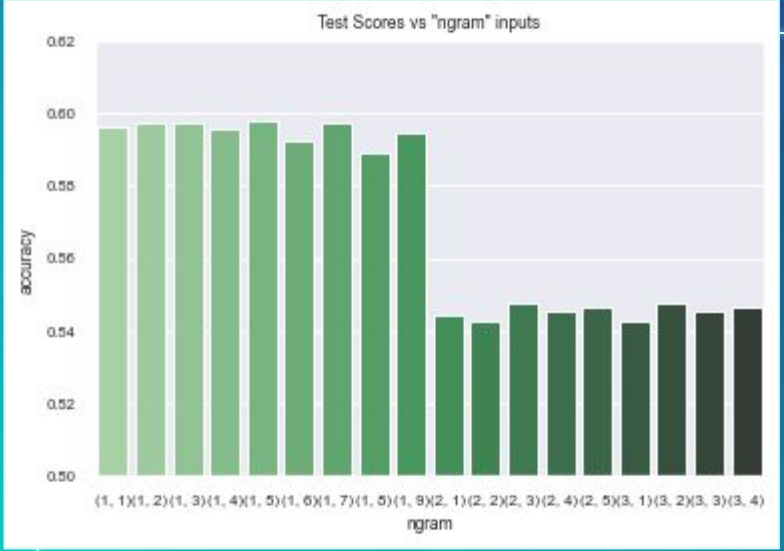
## TOKENS

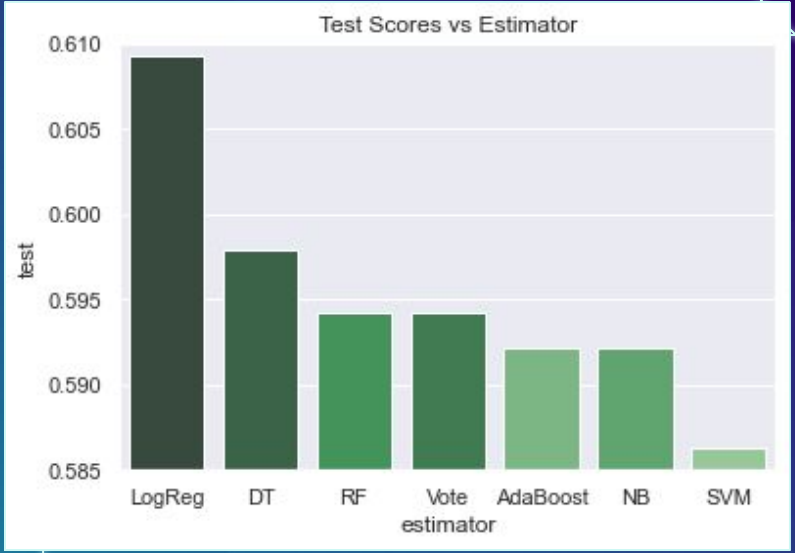
**2**
CountVectorizer, TfidfVectorizer

## 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

# LOGISTICREGRESSIONCV()

5

CV

Cross Validation to ensure reproducibility

200

MAX ITER

To prevent make fitting manageable

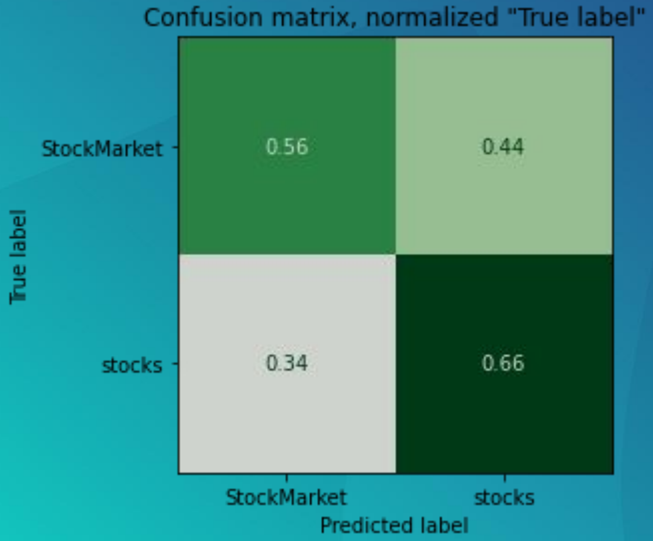
random seed: 42, n_jobs: -1

# GRADING

# MODEL

# LOGISTIC REGRESSION

*“Better than a coin flip”*



Confusion matrix, normalized "True label"

| True label \ Predicted label | StockMarket | stocks |
|---|---|---|
| StockMarket | 0.56 | 0.44 |
| stocks | 0.34 | 0.66 |

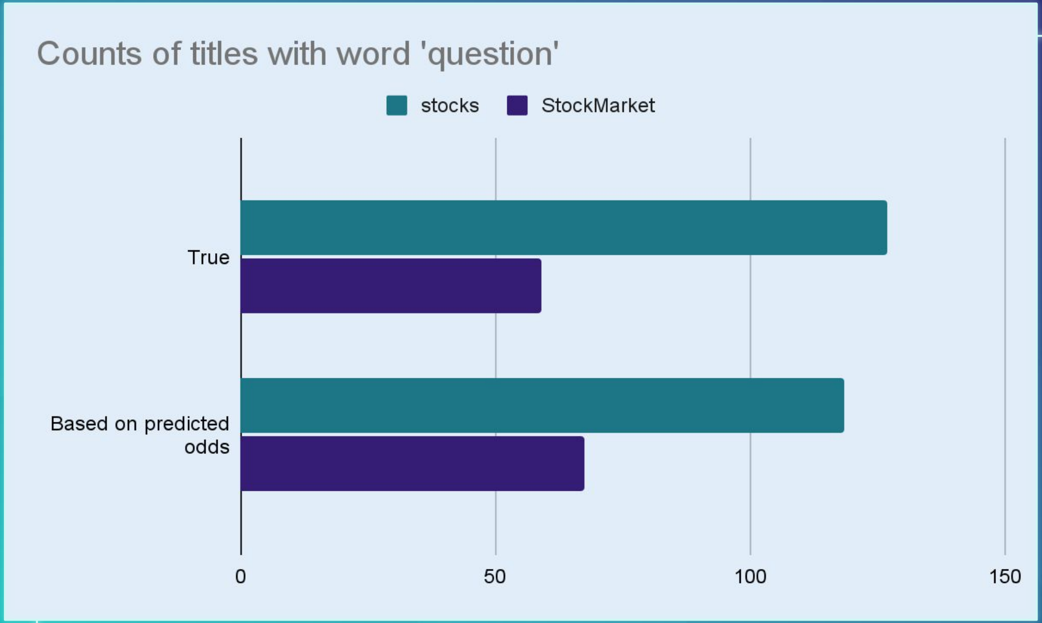
## SCORES

| | |
|---|---|
| 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

in /julian-chang/

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

**Please keep this slide for attribution**