Stonks go up!

Subreddit classification:
r/options vs. r/stocks

GA DSI-20
Samuel Cheah

Problem Statement

What are the most representative text features of posts from each subreddit that will allow us to correctly classify them?

A quick primer

Stocks:

- Indivisible unit of capital representing ownership in a particular
- No expiration

Options:

- A contract that gives the buyer the *option* (hence the name) to purchase 100 underlying shares at a given strike price at expiration.
- Contracts have expiration dates. If the price of the stock goes above your strike price, you get to buy it for cheaper.

Data Cleaning

Scraped the top 1000 posts by month from each subreddit.

After dropping duplicates:

- r/options: 999 posts `0.510938`
- r/stocks: 992 posts `0.489063`

Well balanced classes.

Text Processing

1) Remove URLs

2) Tokenize our text into individual words

3) Remove stop words

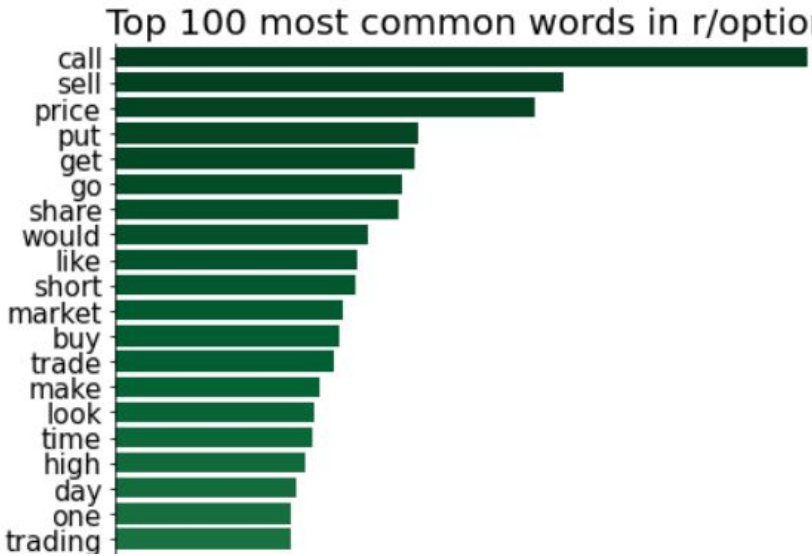
- these are words that appear frequently in the english language and do not provide useful information to us

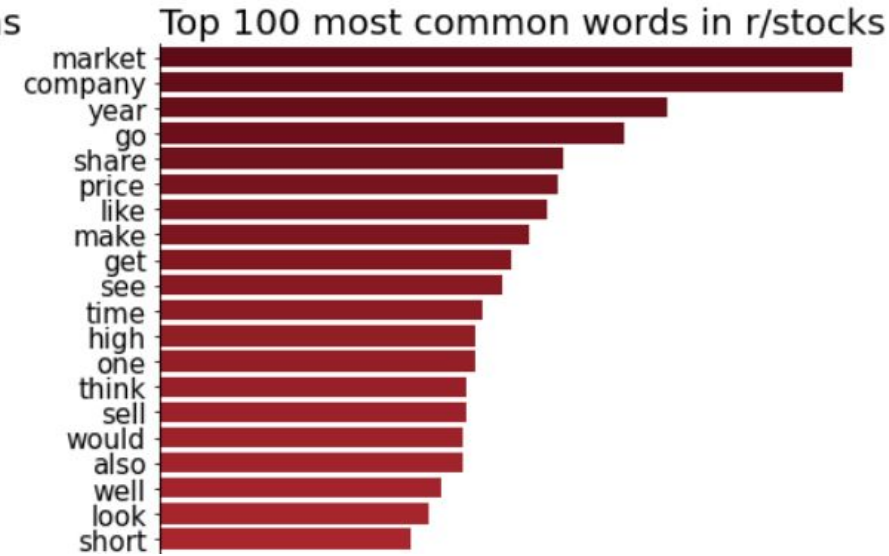
4) Lemmatize our tokens

- this reduces words to their base lemma, allowing us to count different versions of the same word under 1 feature.

5) Rejoin our tokens into a string

EDA


Top 100 most common words in r/options
call
sell
price
put
get
go
share
would
like
short
market
buy
trade
make
look
time
high
day
one
trading


Top 100 most common words in r/stocks
market
company
year
go
share
price
like
make
get
see
time
high
one
think
sell
would
also
well
look
short

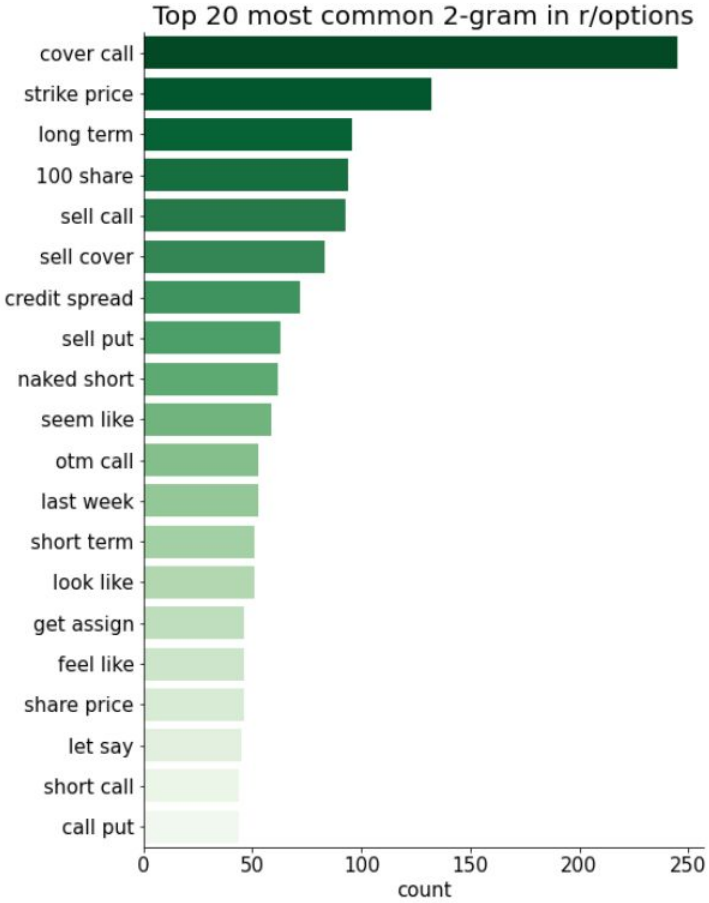
EDA

Options specific technical jargon

- E.g. ‘otm call, ‘strike price’

Focus on trading strategies

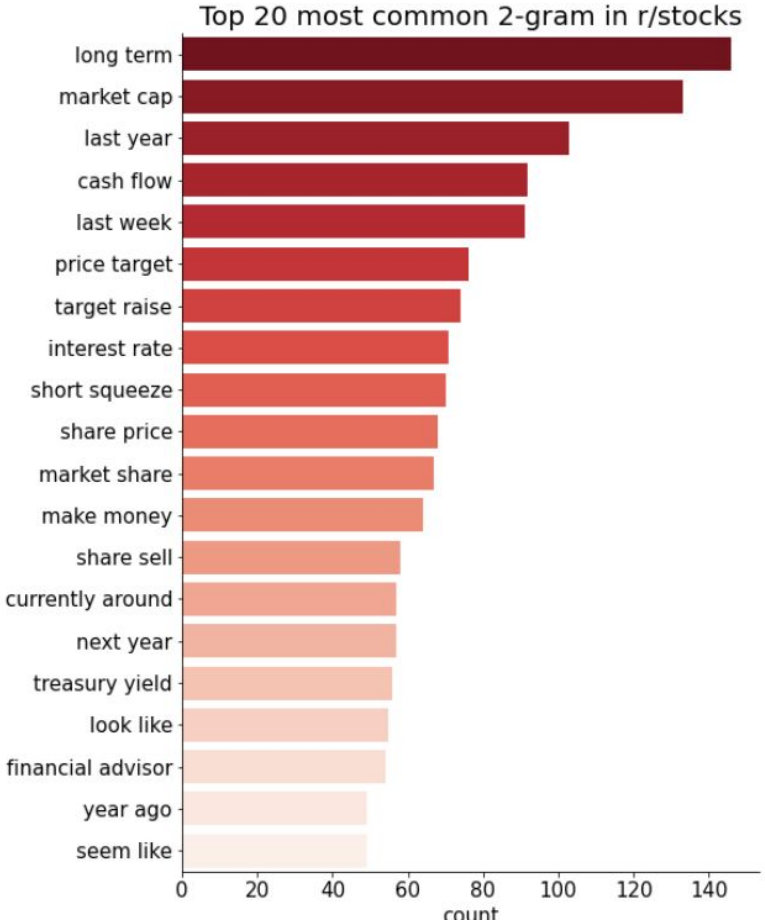
- ‘Cover call’, ‘credit spread’ are names of option trading strategies popular with traders who prefer to *sell* options.



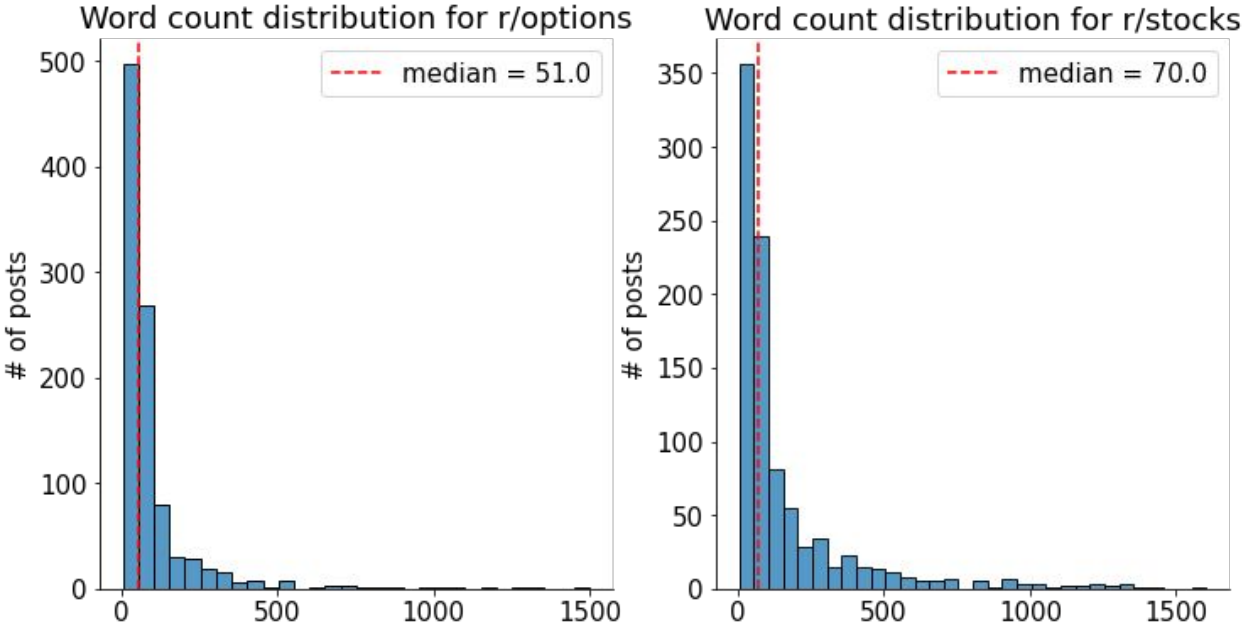
EDA

More focus on market/company fundamentals

- ‘Interest rate’ and ‘treasury yield’ are 2 metrics that have outsized impact on certain classes of stocks, like tech.
- ‘Price target’, ‘cash flow’ are company metrics that traders study to make decisions.



EDA

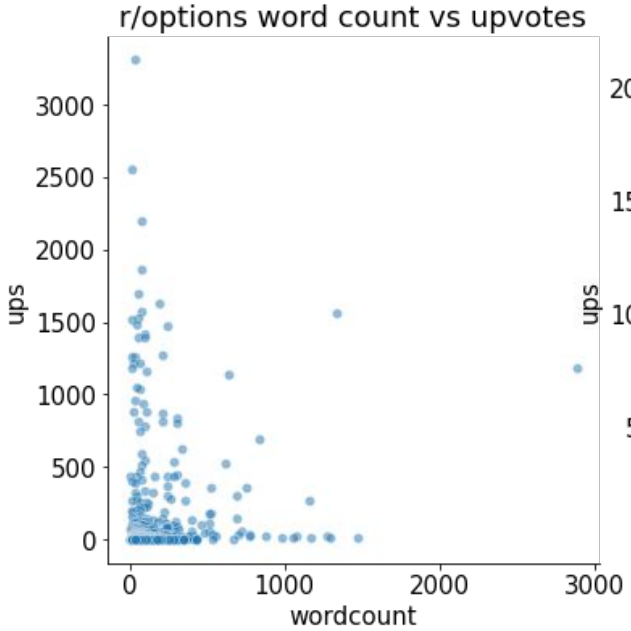

Word count distribution for r/options
median = 51.0
# of posts
0
100
200
300
400
500
0
500
1000
1500
Word count distribution for r/stocks
median = 70.0
# of posts
0
50
100
150
200
250
300
350
0
500
1000
1500

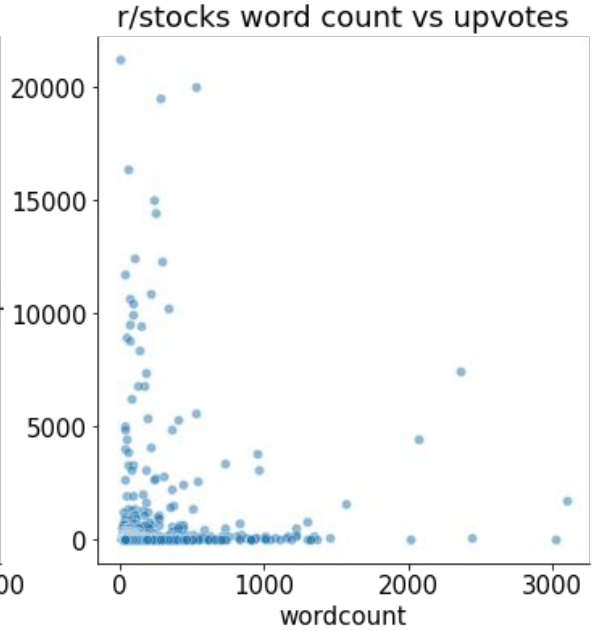
EDA

No clear relationship between word count vs upvotes

EXCEPT

Most popular posts are generally very short.





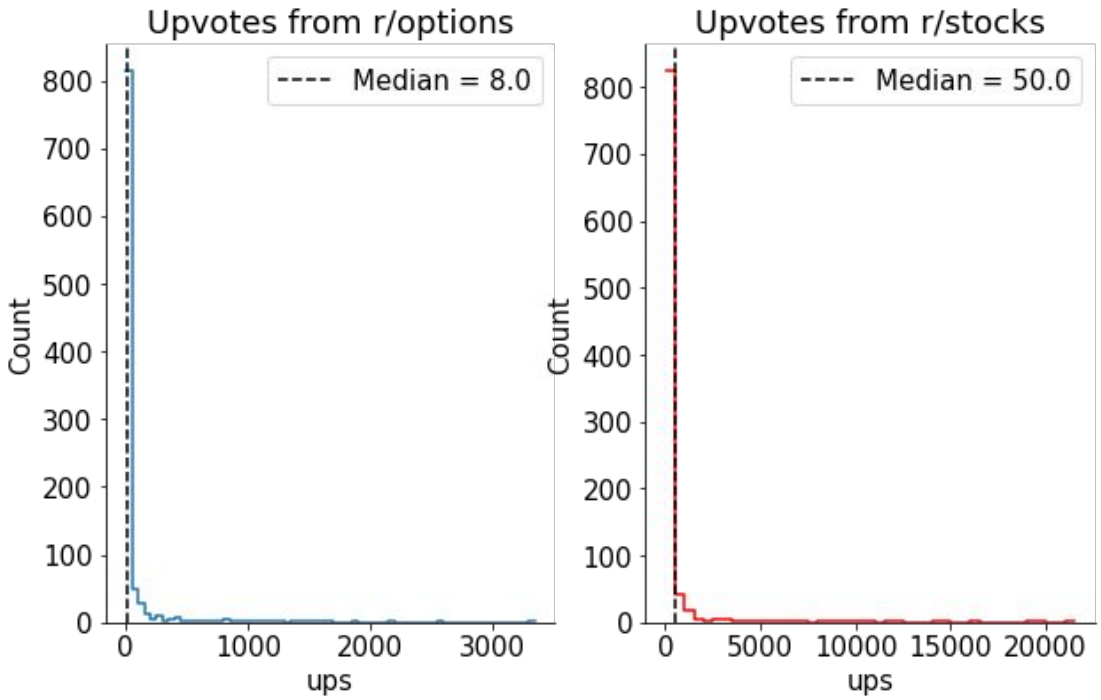
EDA

Subscriber count:

- r/options: 750k
- r/stocks: 2.5mil

r/options vs r/stocks upvote ratio
~ 1:6
r/options vs r/stocks subscriber ratio
~ 1:3

Approx 2x more post engagement in r/stocks than r/options



Model Evaluation

Parametric models did quite well, relatively low overfitting and good ROC AUC and accuracy

Logistic Regression with TfidfVectorizer

```
Training ROC AUC: 0.937
Training accuracy: 0.936
Validation ROC AUC: 0.882
Validation accuracy: 0.881
True positives: 207
False positives: 19
True negatives: 216
False negatives: 38
```

Multinomial NB with TfidfVectorizer

```
Training ROC AUC: 0.939
Training accuracy: 0.939
Validation ROC AUC: 0.869
Validation accuracy: 0.869
True positives: 211
False positives: 29
True negatives: 206
False negatives: 34
```

Model Evaluation

Only 1 tree model gave good results, the rest were very overfit

Best tree model: RandomForest with TfidfVectorizer

```
Training ROC AUC: 0.943
Training accuracy: 0.942
Validation ROC AUC: 0.876
Validation accuracy: 0.875
True positives: 207
False positives: 22
True negatives: 213
False negatives: 38
```

Worst tree model: AdaBoost with TfidfVectorizer

```
Training ROC AUC: 1.000
Training accuracy: 1.000
Validation ROC AUC: 0.864
Validation accuracy: 0.863
True positives: 199
False positives: 20
True negatives: 215
False negatives: 46
```

Final model: Logistic Regression + TVec

Our 3 best models were Logistic Regression, Multinomial Naive Bayes, and Random Forest

Vectorizer	Estimator	Train accuracy	Validation accuracy	Train ROC AUC	Validation ROC AUC
tvec	Logistic Regression	0.936	0.881	0.937	0.882
tvec	Multinomial NB	0.939	0.869	0.939	0.869
tvec	RandomForest	0.942	0.875	0.943	0.876

Logistic Regression + Tvec: **Least overfit, highest scores.**

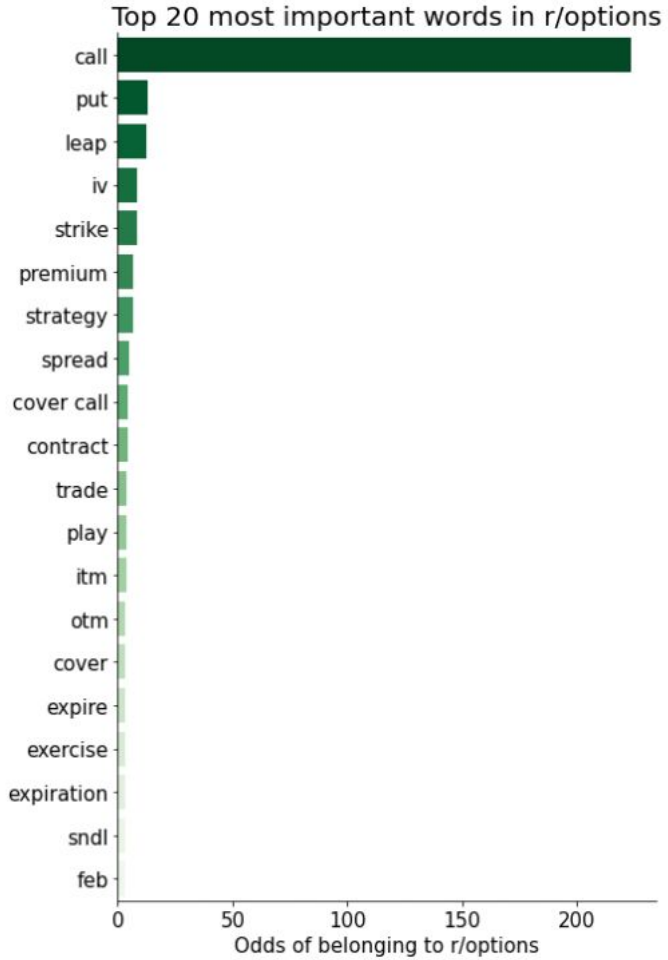
Analysis of results
- r/options

‘Call’ has outsized importance in r/options.

- Is a legitimate feature however, so we leave it in.

Option specific terms and concepts heavily featured here

- E.g. ‘cover call’, ‘expiration’, ‘iv’, ‘strike’.



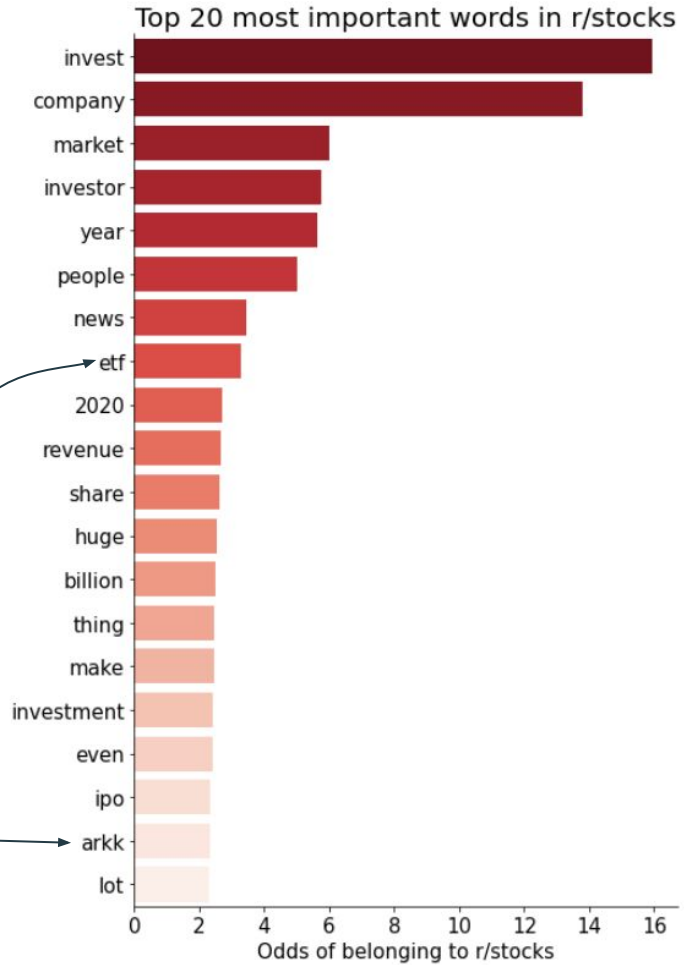
Analysis of results
- r/stocks

More emphasis on high level, fundamental analysis

- Terms like ‘invest’, ‘revenue’, ‘market’, ‘company’ show focus on fundamental analysis rather than technical.
- Less emphasis on technical analysis like trade volume or usage of indicators e.g. RSI, 50 SMA.

Stock traders like discussing Exchange Traded Funds, (‘etf’).

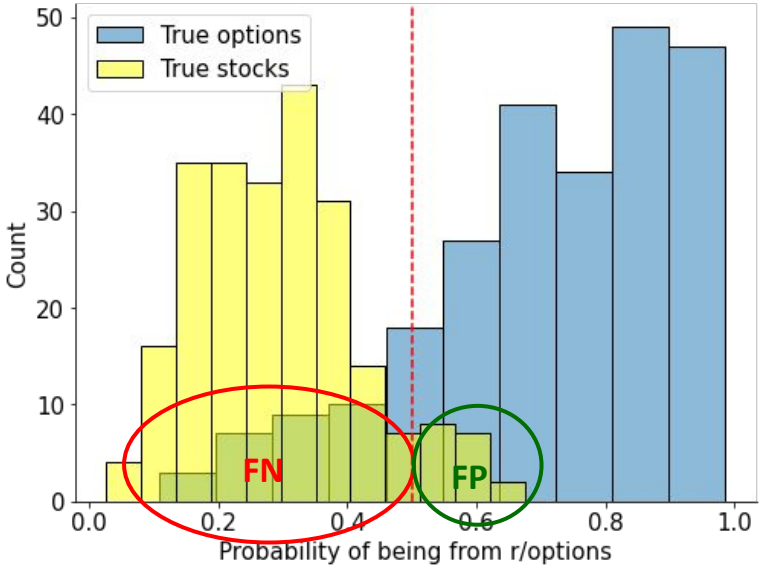
- ARKK is one of the hottest ETFs on the market now.

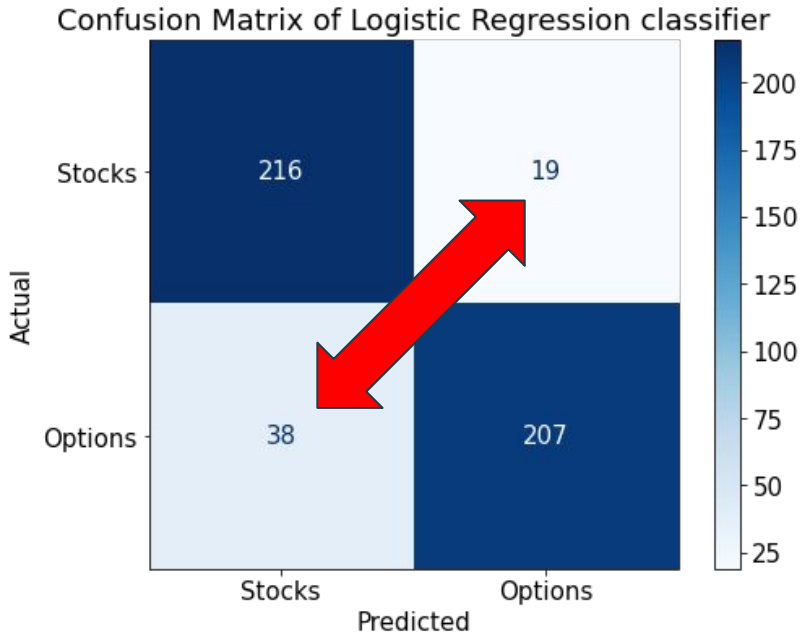


Misclassified posts

More False Negatives than Positives.

Classifier worse at classifying r/options posts.





Misclassified posts - False Negative

r/options posts misclassified as r/stocks:

Due diligence post

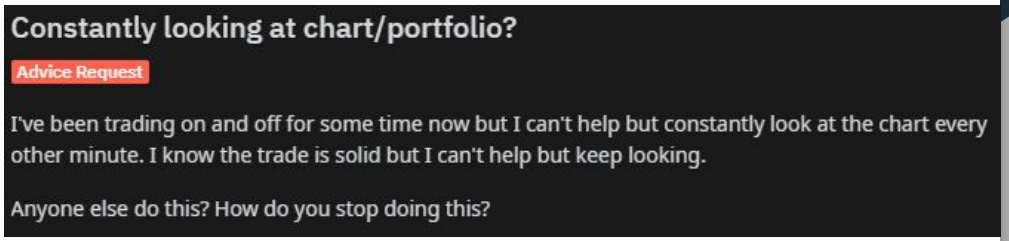
- DD refers to research done on a company or market sector.
- Poster usually presents his trading thesis
- Lots of fundamental analysis

Very little mention of anything to do with options

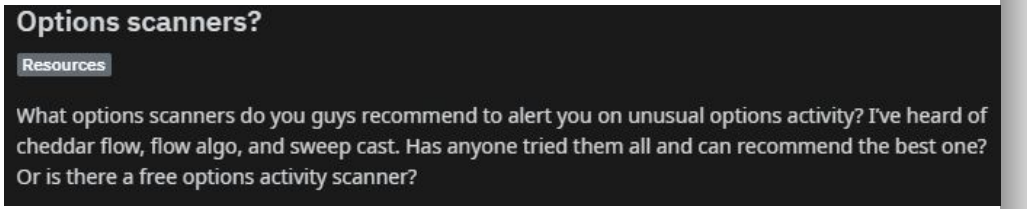
Misclassified posts - False Positive

r/stocks posts misclassified as r/options:

Generic posts asking for advice.



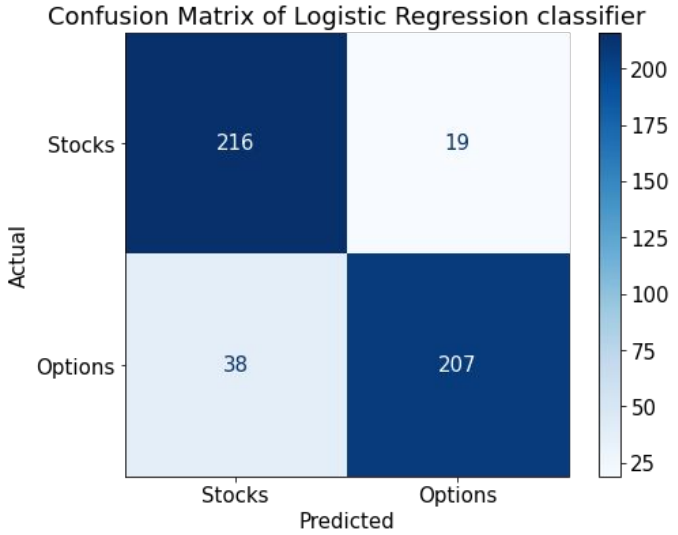
Wrong subreddit!!! Should have been posted to r/options instead.



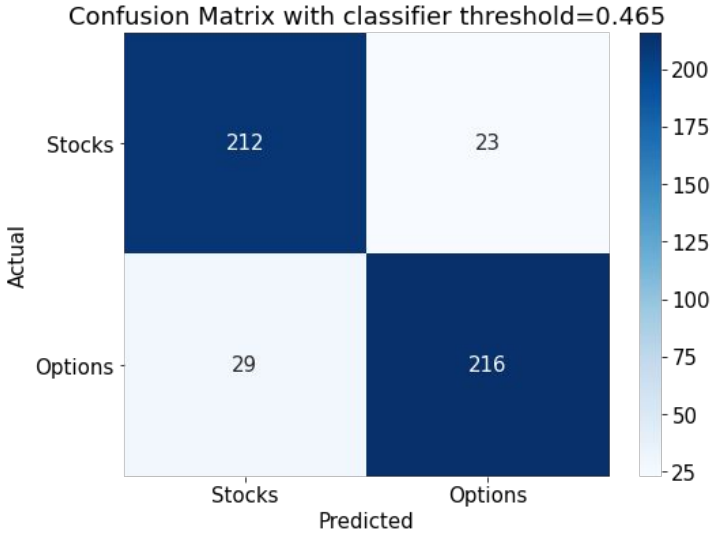
Improvements to model

Removed 3 additional words that came up frequently in misclassified posts

- ‘Put’, ‘short’, ‘premium’ added as stopwords



Training ROC AUC: 0.937
Training accuracy: 0.936
Validation ROC AUC: 0.882
Validation accuracy: 0.881



Training ROC AUC: 0.936
Training accuracy: 0.936
Validation ROC AUC: 0.890
Validation accuracy: 0.890

Recommendations

1) Scrape data every month to increase size of training corpus.

2) Use name entity recognition (spaCy) and sentiment analysis (VADER) to analyze retail sentiment surrounding specific companies and popularity within subreddit.

	feature	coef	
2479	sndl	3.168640	More likely to appear in r/options
2533	spy	2.708709	
210	amc	2.013681	
2736	tlry	1.181805	
219	amzn	0.944795	
126	aapl	0.683463	More likely to appear in r/stocks
1225	gme	0.530063	
380	blackberry	0.526046	