Project 3: Web APIs & Classification

Group 4
Fionna
Jia Chi
Shu Jun
Zhi Yong

Overview

1. Problem Statement
2. Web Scraping from Reddit
3. Data Cleaning & Feature Engineering
4. Exploratory Data Analysis
5. Modelling & Validation
6. Key Findings, Conclusions & Recommendations

Problem Statement - Investment

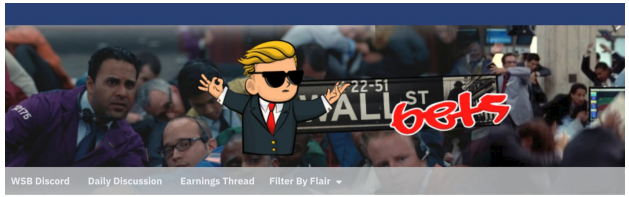
Online investment forums (such as Reddit) can aid us in choosing the right stock for investment, and for assessing the “market sentiment” of different stocks types.

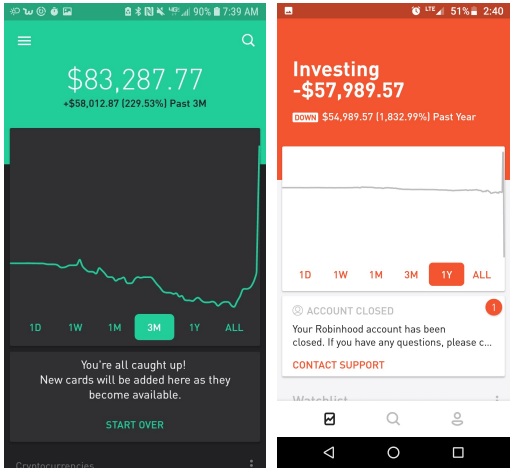
- r/wallstreetbets (high-risk/high-reward, almost reckless)
- r/investing (safer forms of investment portfolios)

However, blindly taking financial advice without consideration of the background origins can be very dangerous financially.

Different Profiles between the Subreddits

r/wallstreetbets (745k members)





"high-risk, high-reward"

r/investing (845k members)

Posts | Contact Moderators

Posted by u/benne362 14 hours ago
34
First time investing outside of 401(k)
I currently max out my 401(k) and have a good emergency funds set aside. I dint have any debt outside of my mortgage, so I'm looking to start investing. I'm not looking to day trade, just open an account somewhere and take on some mutual funds (not individual stocks).

Posted by u/AntiCirclejerking 16 hours ago
92
Can you make a living off of investing in S&P?
I'm having trouble understanding how someone can gain cash flow from stocks. If you have $250k and invest it into S&P that's great, now what? Your money will historically go up but when do you actually sell and cash out? How can you acquire passive income from this and make a living off of it? Or is the goal to just hold your money in the

Posted by u/Jones743 17 hours ago
82
Gold and silver bars
I've been looking into several threads and it always seems to go back and forth. When buying smaller bars 10oz, 5oz, 1oz, the bar comes in a plastic container (wrap?). Taking it out of the plastic makes it easier to store, but does it affect the value? The bars still have

"low-risk, stable returns"

Problem Statement - Investment

The aim of this project is to build a NLP classifier that will help us **distinguish the message posts between these 2 popular subreddits**, so that we can make **informed investment decisions based on our desired investment objectives and risk appetites.**

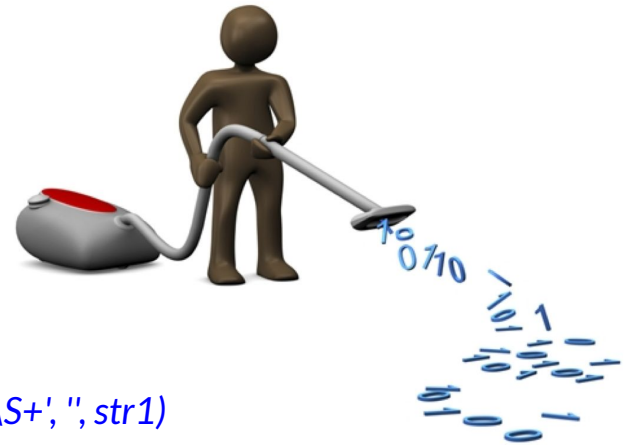
Web Scraping from Reddit

Reddit API

- Set anonymous username to gain access
- Set random sleep duration to look more ‘natural’
- Save data to .csv file
- Very few posts scraped for r/wallstreetbets (~400 posts)
    - Use Pushshift.io API (access to database for older posts)
    - https://github.com/dmarx/psaw

Data Cleaning

1. Extract text using Beautifulsoup
2. Tokenization + Remove punctuation
3. Remove stop words
4. Lemmatization
5. Remove hyperlinks using Regex e.g. *re.sub(r'http\S+', '', str1)*
6. Identify and Remove Duplicates



Identify and Remove Duplicates

	author	created_utc	full_link	selftext	subreddit	title
8979	SonOfAntonopolis	1565217192	https://www.reddit.com/r/investing/comments/cn...	https://youtu.be/ICL28qm2634	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
8152	cryptosplash	1565988404	https://www.reddit.com/r/investing/comments/cr...	https://youtu.be/ICL28qm2634	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
8028	SuperCryptoX	1566136202	https://www.reddit.com/r/investing/comments/cs...	https://youtu.be/ICL28qm2634	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
7313	bountyzzzz	1566896056	https://www.reddit.com/r/investing/comments/cw...	[removed]	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
7257	AdamR3333	1566939573	https://www.reddit.com/r/investing/comments/cw...	https://youtu.be/ICL28qm2634	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
9415	VitaliksUnicorn	1564779882	https://www.reddit.com/r/investing/comments/cl...	https://youtu.be/ICL28qm2634	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
9315	BitcoinBoss1234	1564958294	https://www.reddit.com/r/investing/comments/cm...	https://youtu.be/ICL28qm2634	investing	$25 Free Bitcoin! No Purchase Required (U.S.A ...
17399	trader992	1554912183	https://www.reddit.com/r/investing/comments/bb...	\n\nwow! what a trading result must watch \n...	investing	$485 Turned into Over $244,000 in Just 75 Days...
17394	AlternativeOrchid	1554921576	https://www.reddit.com/r/investing/comments/bb...	[http://www.fxchildsplaysignals.com/](http://...	investing	$485 Turned into Over $244,000 in Just 75 Days...

Feature Engineering

Decrease null values and increase useful information

```
wsb_posts_push.head(3)    # Checking the format for r/wallstreetbets
```

	author	created_utc	full_link	selftext	subreddit	title
0	Gotcha_Scumbag	1575309062	https://www.reddit.com/r/wallstreetbets/commen...		wallstreetbets	More free money from Robinhood. Just open an a...
1	Orgasimo	1575308877	https://www.reddit.com/r/wallstreetbets/commen...		wallstreetbets	u/RobinhoodTeam tries to manage r/WallStreetBets
2	Jaseiker	1575308753	https://www.reddit.com/r/wallstreetbets/commen...	Between the press coverage of being on the fro...	wallstreetbets	WSBs has become a valuable brand, is anyone mo...

Concatenate ‘Title’ and ’selftext’ as X input

Exploratory Data Analysis

Use count vectorization to get number of counts for each word - Word Cloud

	words	counts
1694	stock	13819
1082	market	10561
1986	year	9491
357	company	8112
1974	would	7760
1016	like	7409
1147	money	7122
739	fund	6365
1792	time	6012
255	buy	5701
932	investing	5457
1590	share	5359
764	get	5355
1226	one	4868
973	know	4680

Modelling

1. Logistic Regression/Count Vectorizer
2. Logistic Regression/Tf-idf Vectorizer
3. Logistic Regression/Count Vectorizer(GridCV)
4. Logistic Regression/Tf-idf Vectorizer(GridCV)
5. Multinomial Naive-Bayes/Count Vectorizer (GridCV)

Model Metrics (Comparison)

Model	Accuracy (Train)	Accuracy (Test)	ROC-AUC Score
1. Logistic Regression / Count Vectorizer	0.995	0.837	0.898
2. Logistic Regression / Tf-idf Vectorizer	0.960	0.875	0.925
3. Logistic Regression / Count Vectorizer (Grid CV)	0.998	0.831	0.903
4. Logistic Regression / Tf-idf Vectorizer (Grid CV)	0.963	0.872	0.928
5. Multinomial Naive-Bayes / Count Vectorizer (Grid CV)	0.910	0.853	0.918

Model Selected

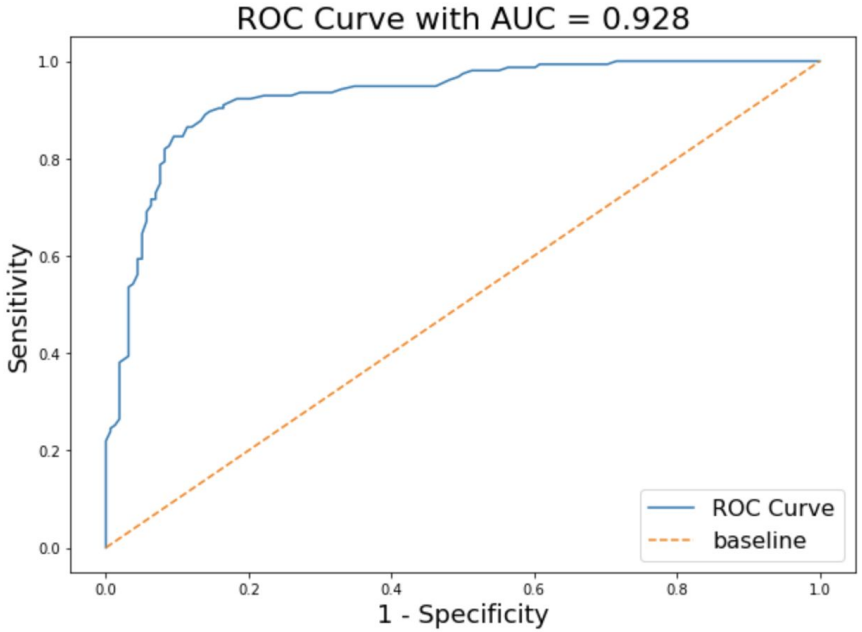
Logistic Regression / Tf-idf Vectorizer (Grid CV)

Best Parameters Found:

```
{'tvec__max_df': 0.9,
 'tvec__max_features': 3500,
 'tvec__min_df': 2,
 'tvec__ngram_range': (1, 2)}
```

Model Selected

Logistic Regression / Tf-idf Vectorizer (Grid CV)



Model Selected

Logistic Regression / Tf-idf Vectorizer (Grid CV)

	pred wallstreetbets	pred investing
actual wallstreetbets	130	28
actual investing	18	137

```
# Number of correct positive predictions divided by total number of positives

sens = tp / (tp + fn)

print(f'Sensitivity: {round(sens,4)}')
```

```
Sensitivity: 0.8839
```

```
# Number of correct negative predictions divided by the total number of negatives

spec = tn / (tn + fp)

print(f'Specificity: {round(spec,4)}')
```

```
Specificity: 0.8228
```

Test set of 313 posts

With A Larger DataSet

Logistic Regression / Tf-idf Vectorizer (Grid CV)

	pred r/inv	pred r/wsb
actual r/inv	6425	998
actual r/wsb	949	6357

```
sensitivity = round(tp/(tp+fn),4)
print(f"Our model's Sensitivity is: {sensitivity}")
print(f"Out of 100 r/wsb posts, our model is able to correctly classify {int(sensitivity*100)} of
them.")
```

```
Our model's Sensitivity is: 0.8701
Out of 100 r/wsb posts, our model is able to correctly classify 87 of them.
```

```
specificity = round(tn/(tn+fp),4)
print(f"Our model's Specificity is: {specificity}")
print(f"Out of 100 r/inv posts, our model is able to correctly classify {int(specificity*100)} of
them.")
```

```
Our model's Specificity is: 0.8656
Out of 100 r/inv posts, our model is able to correctly classify 86 of them.
```

Test set of 14729 posts

Key Findings

1. Key words (slang) between the two subreddits enables robust message classification.

2. Allows for additional insights into the subreddit's demographic and habits.

3. Insights into their stock trading preferences and their trading strategy/plans.

4. NLP model however fails to distinguish the message context, or if the message is too "general" in nature.

Top 20 Most Frequently Occuring Words

1. Key words (slang):
   - tendies
   - yolo
   - autists
   - dd

   } r/wsb

2. Trading habit:
   - call/put (options) (r/wsb)
   - stocks/indexes (r/inv)
   - bonds/ira (r/inv)

	words	log_coeff	coeffs
5923	wsb	6.25217	519.138965
5245	tendies	5.96053	387.815070
5978	yolo	5.74414	312.356172
1279	dd	5.62978	278.600804
4735	shit	5.47178	237.882328
365	autists	5.35944	212.605526
363	autist	4.59696	99.182817
753	call	4.45505	86.060565
4418	retard	4.3986	81.336866
4153	put	4.18376	65.612249
4419	retarded	4.1349	62.483159
2108	fuck	4.00225	54.721214
362	autism	3.68267	39.752410
364	autistic	3.63457	37.885410
2112	fucking	3.57059	35.537492
4459	rh	3.5456	34.660308
3177	man	3.45937	31.796806
628	boy	3.27764	26.513259
2350	guh	3.06566	21.448520
2171	gang	2.94904	19.087566

r/wallstreetbets

	words	log_coeff	coeffs
2709	investing	-6.23027	0.001969
2722	investment	-5.5934	0.003722
943	cnbc	-4.75161	0.008638
2737	investor	-4.53362	0.010742
1441	dividend	-4.07222	0.017039
2696	invest	-4.05391	0.017354
1714	etf	-3.99855	0.018342
5648	vanguard	-3.93814	0.019484
5857	wondering	-3.79344	0.022518
2122	fund	-3.47373	0.031001
2588	index	-3.37668	0.034161
5876	would	-3.31571	0.036308
4508	roth	-3.25194	0.038699
4985	stock	-3.18919	0.041205
296	article	-3.15264	0.042739
590	bond	-3.11024	0.044590
451	bear market	-3.10958	0.044620
5248	term	-3.04294	0.047695
2758	ira	-3.03653	0.048001
998	company	-2.9388	0.052929

r/investing

Wrongly Classified Posts

r/wallstreetbets: wrongly classified as r/investing

```
#1232 gain almost year watch buffet

#18054 investing chinese money exodus many wealthy chinese people transferring huge amount money c
hina primarily canada u recommendation take advantage situation investing certain bank american ca
nadian brand appeal wealthy chinese people
```

r/investing: wrongly classified as r/wallstreetbets

```
#46659 ok new reddit investing term like tendie autist never heard

#36057 good sense stock go bad buy sell dozen time make little profit
```

Two possible reasons for wrong classification:

- keyword spamming
- “general” sounding posts

Top 10 Most Frequently Occuring Stock Tickers

SPX500 Stock Tickers

1. r/wallstreetbets:
   - Activision Blizzard
   - Ulta Beauty
   - General Electric
   - AMD
   - Disney
2. r/investing:
   - Alphabet
   - Costco
   - Morgan Stanley Capital Investment
   - Proctor & Gamble
   - Electronic Arts

	words	log_coeff	coeffs
1279	dd	5.62978	278.600804
5272	tgt	2.362	10.612130
3611	nov	1.92463	6.852625
345	atvi	1.41981	4.136332
5554	ulta	1.41429	4.113557
2186	ge	1.4081	4.088194
174	amd	1.11981	3.064279
5799	well	0.879529	2.409765
1402	dis	0.840638	2.317845
3575	nflx	0.746436	2.109469

r/wallstreetbets

	words	log_coeff	coeffs
2289	googl	-1.94169	0.143462
1141	cost	-1.54536	0.213234
3458	msci	-1.17558	0.308640
3876	pg	-1.15719	0.314369
1527	ea	-1.01486	0.362452
1283	de	-0.971526	0.378505
2517	ibm	-0.885528	0.412496
2842	khc	-0.819902	0.440475
3621	nvda	-0.680445	0.506392
5604	ups	-0.670578	0.511413

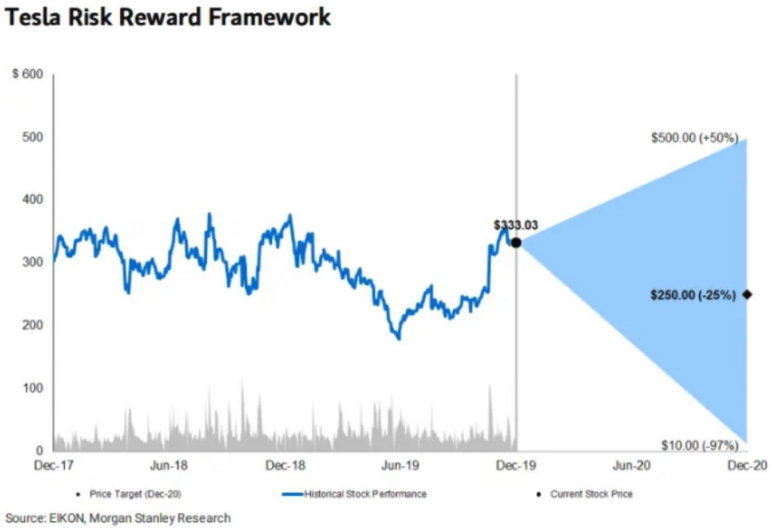
r/investing

“Title without message body” posts

Posted by u/Wyxuan 9 minutes ago

Real analyst report from Morgan Stanley

Stocks



5 Comments Share Save Hide Report 100% Upvoted

Meaning: MS analysts do not know what they are doing

r/wallstreetbets · Posted by u/WSBConsensus a useful lad 23 minutes ago

Spotted in the wild.

Discussion

13 Comments Share Save Hide Report 96% Upvoted

Meaning: People are buying TSLA’s cybertruck

Recommendations

1. Improving the removal of “noise words”:
    a. HTML, webpage links and image artifacts.
2. Determining how current Reddit posts forecast future trends
    a. How real-time posts change based on current market trends.
    b. Understanding the market sentiment can be helpful in predicting future trends.
3. Capturing of image context
    a. Reddit post with only a title without message body.
    b. Meme posts.

Thanks!