Reddit Classification Project: Analyzing r/bitcoin and r/wallstreetbets

Dennis Stoliaryk - Data Scientist

What is Reddit and the subreddits we are interested in?

Reddit is the most popular internet forum, called "the frontpage of the internet."

Individuals can post anonymously on virtually any topic in subreddits (dedicated forum to a certain subject on reddit) by registering a username and submitting a post/comment.

r/bitcoin is a subreddit dedicated to discussing Bitcoin, the "currency of the internet."

r/wallstreetbets is a subreddit dedicated to postings of risky investments from individuals who are "yoloing" their life savings.

Problem Statement

Can we accurately predict a subreddit from a reddit post's title using classification modeling?

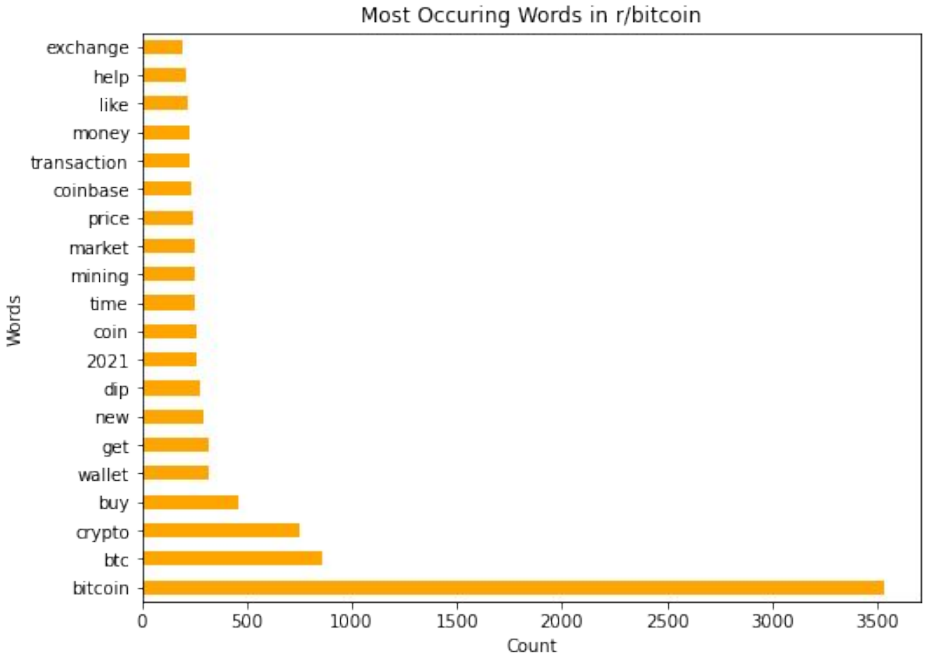
Data Collection and Gathering

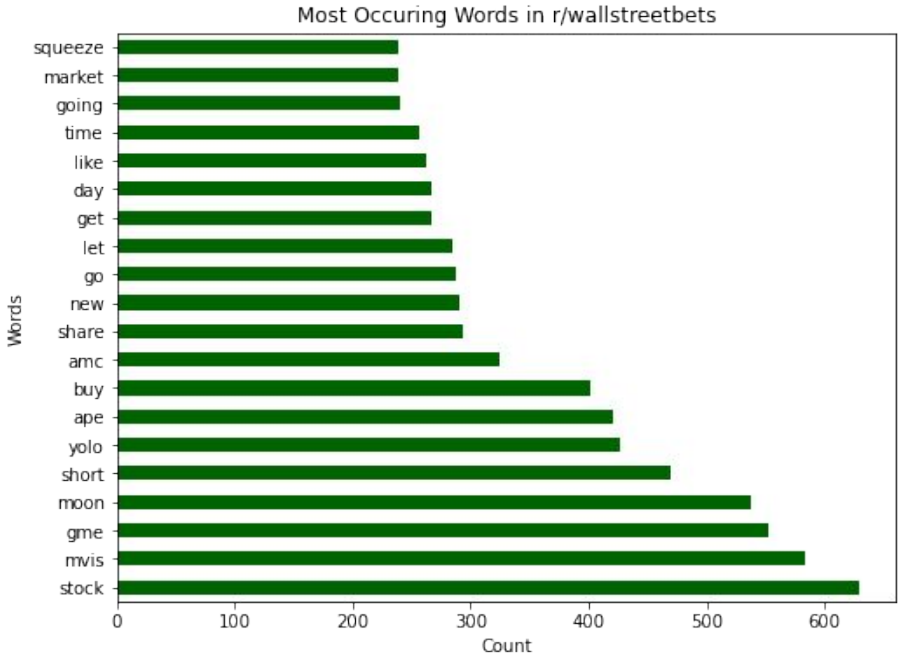
Used Pushshift's API to request and pull data from r/bitcoin and r/wallstreetbets.

Data Cleaning and EDA

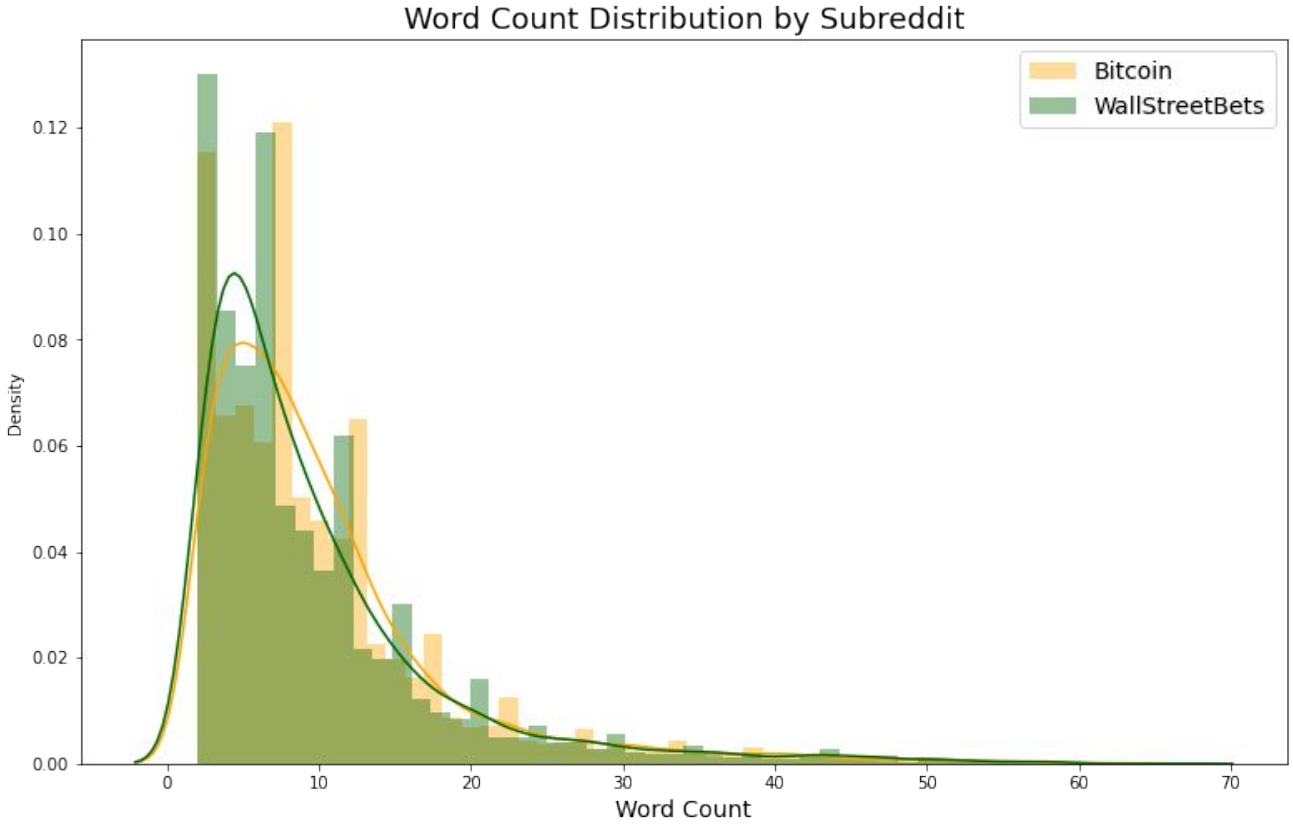
- Removed duplicate titles
- Dropped missing titles
- Removed titles that were one word long
- Tokenized/Lemmatized our titles for better analysis.

EDA


Most Occuring Words in r/bitcoin
Words
exchange
help
like
money
transaction
coinbase
price
market
mining
time
coin
2021
dip
new
get
wallet
buy
crypto
btc
bitcoin
0
500
1000
1500
2000
2500
3000
3500
Count


Most Occuring Words in r/wallstreetbets
Words
squeeze
market
going
time
like
day
get
let
go
new
share
amc
buy
ape
yolo
short
moon
gme
nvis
stock
0
100
200
300
400
500
600
Count

EDA


Word Count Distribution by Subreddit
Bitcoin
WallStreetBets
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
60
70
Word Count

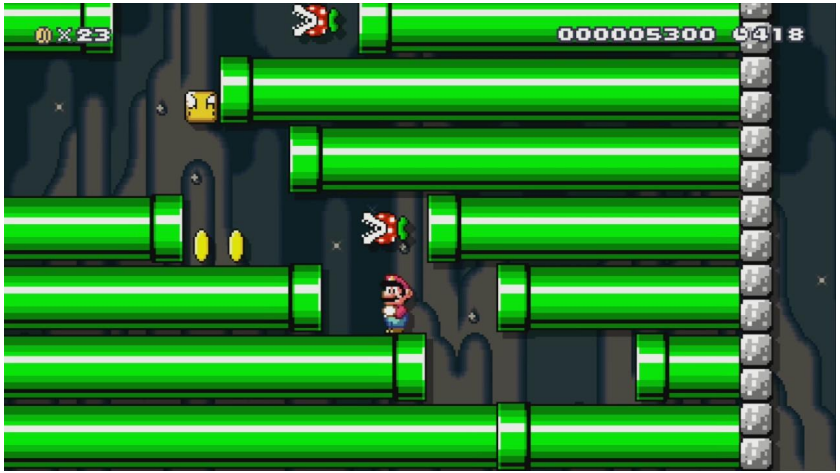
Modeling

Baseline Score of ~50%

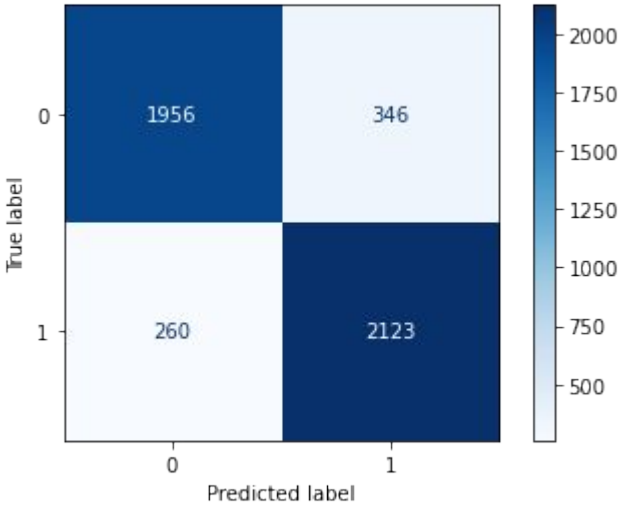
Pipeline using CountVectorizer and Naive Bayes (Multinomial NB)

GridSearching

Training Score: 96%
Testing Score: 87%



Modeling Performance (Confusion Matrix)



	precision	recall	f1-score	support
0	0.8827	0.8497	0.8659	2302
1	0.8599	0.8909	0.8751	2383
accuracy			0.8707	4685
macro avg	0.8713	0.8703	0.8705	4685
weighted avg	0.8711	0.8707	0.8706	4685

Top 20 Most Predictive Words

	coef	words
8043	-3.952265	bitcoin
10399	-5.374243	btc
15841	-5.493590	crypto
11084	-5.992167	buy
57860	-6.424066	wallet
18084	-6.455814	dip
37849	-6.479125	new
35854	-6.542407	mining
1305	-6.547445	2021
54110	-6.552508	time
13762	-6.583438	coinbase
13589	-6.593965	coin
34573	-6.609965	market
42530	-6.665228	price
36370	-6.670926	money
55174	-6.699913	transaction
32355	-6.711748	like
23070	-6.825076	free
26961	-6.825076	help
14	-6.831765	000

	coef	words
90565	-6.037630	stock
65982	-6.132519	mvis
43379	-6.153572	gme
65099	-6.222059	moon
86460	-6.348742	short
8702	-6.435753	ape
19913	-6.461152	buy
106719	-6.487212	yolo
7562	-6.703776	amc
67309	-6.789298	new
85791	-6.807317	share
56601	-6.821047	let
57302	-6.877935	like
30197	-6.943449	day
95820	-6.975198	time
61342	-7.024795	market
43935	-7.036159	going
96455	-7.041889	today
89491	-7.071046	squeeze
44498	-7.144699	good

Conclusions/Further Study

We can accurately predict which subreddit a title is from with an accuracy over 85%!

r/wallstreetbets predicting words are stock, GME, MVIS, AMC, yolo, ape, etc.

r/bitcoin predicting words are bitcoin, btc, crypto, wallet, dip, mining, hodl.

For further research, I would recommend gathering entire data sets from subreddits.

Use different modeling techniques as well to get the best accuracy possible.

Symbol	Actions	Last Price $	Change $	Change %	Qty #	Price Paid $	Day's Gain $	Total Gain $	Total Gain %	Value $
GME		194.50	56.76	41.21%	100,000	26.7986	5,676,000.00	16,770,138.84	625.78%	19,450,000.00
GME Apr 16 '21 $12 Call		177.25	56.35	45.16%	500	0.20	2,817,500.00*	9,045,991.80	88,183.03%	9,056,250.00
Cash Total Transfer money										$11,882,936.46
Total						$2,690,119.36	$8,493,500.00	$25,816,130.64	959.66%	$40,389,186.46

Positions

Quick Trade

CALL ALERT - IMMEDIATE ACTION REQUIRED

Symbol	Mark	P/L YTD ▲	Cost
AAPL ITM APPLE INC COM	131.50	($184,987.51)	$131.3756
PLTR ITM PALANTIR TEC...	23.00	($3,851.00)	$24.684
VIX CBOE MARKET VOL...	20.45	($60.85)	—
F ITM FORD MOTOR...	11.53	$53.00	—
Overall Totals:	—	($188,846.36)	—

P/L Day: ●●●●●●●●●

P/L Open: ●●●●●●●●●

Net Liq: ●●●●●●●●●

Available $: ●●●●●●●●●

Position Equity: ●●●●●●●●●

Watchlist | Positions | Orders | Alerts | More

Thank you! Any Questions?