Riding the Waves of Crypto and Stocks: NLP-Powered Voyage through Reddit's Investment Chatter

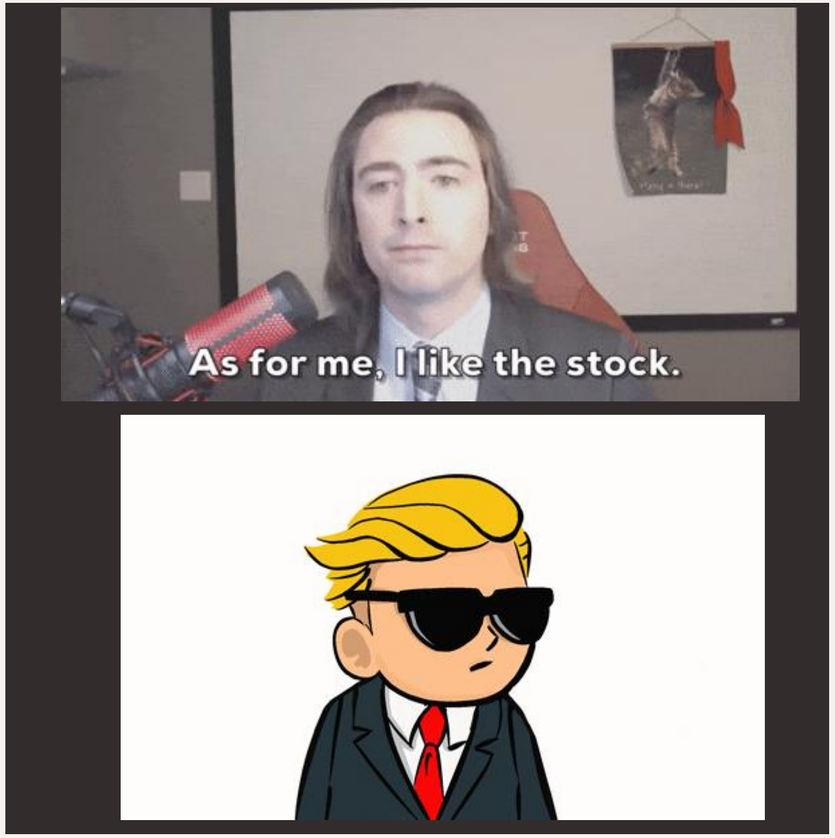
Develop a binary NLP classifier to distinguish between crypto and stock posts found on the r/wallstreetbets and r/CryptoMoonShots subreddits. This will help users identify investment opportunities in both markets and potentially inform trading decisions based on which assets are popular/rising. #ShortSqueeze #TotheMoon

r/wallstreetbets

- **Meme Stonks & Retail Rebellion:** WallStreetBets rose to fame during the pandemic with the GameStop saga. Users coordinated massive stock purchases, popularizing the term "meme stocks" and challenging Wall Street institutions. This "casino-style" approach to the market has become a point of legal concern.

- **RoaringKitty (Keith Gill):** popularized the GameStop short squeeze in 2021, raising questions about retail investor influence and sparking a trial that could define legal boundaries for online financial communication.

- **Financial Lingo Goes Viral:** WSB's crass humor and unique financial slang ("diamond hands," "to the moon") became mainstream conversation in financial news.

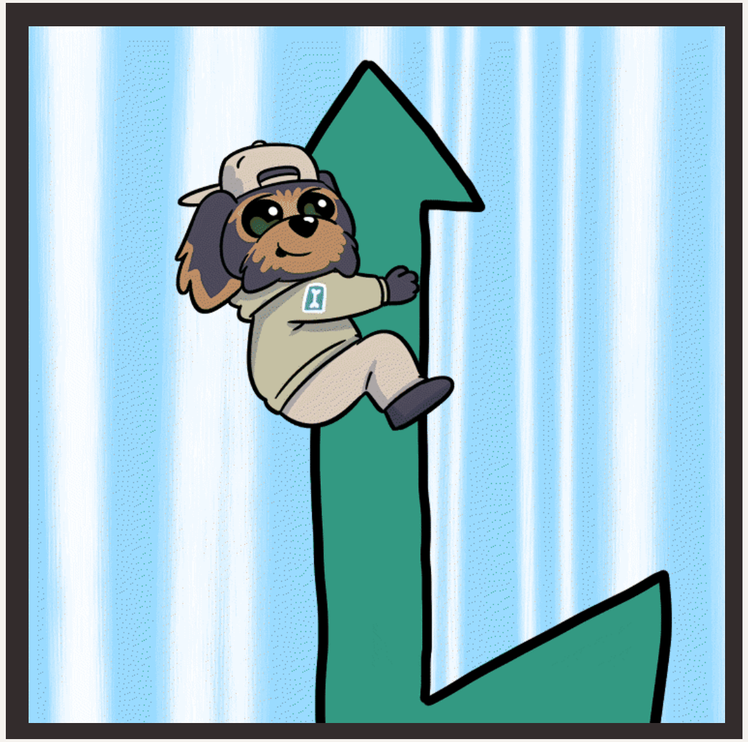


r/CryptoMoonShots

- **Community Shilling & High Risk:** The subreddit thrives on user recommendations and promotions (shilling) of unproven cryptocurrencies. This can lead to high-risk investments and frequent scams.

- **"To the Moon" & Memes:** Similar to WallStreetBets, CryptoMoonShots uses memes and internet slang ("to the moon," rocket emojis) to express excitement about crypto gains.

- **Focus on Price & Short-Term Gains:** Discussions often center around short-term price movements and potential profits, rather than the underlying technology or long-term viability of crypto projects.



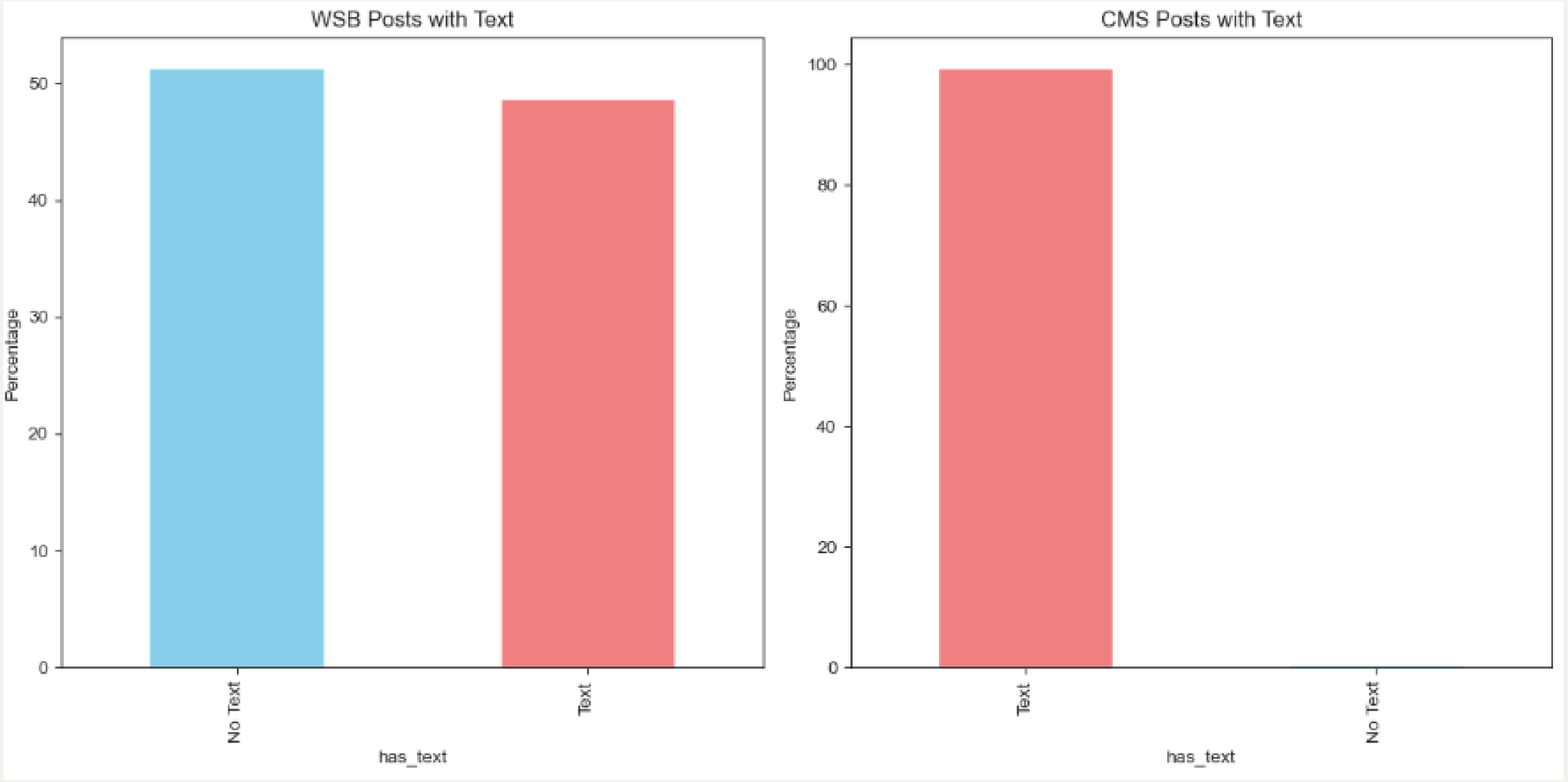
Objectives

1.Gather and prepare data from the wallstreetbets and CryptoMoonShots subreddits using PRAW (Python Reddit API Wrapper). Extract from top, new and controversial sections of each subreddit.

2.Preprocess the text data by cleaning, tokenizing, and vectorizing the posts for NLP analysis.

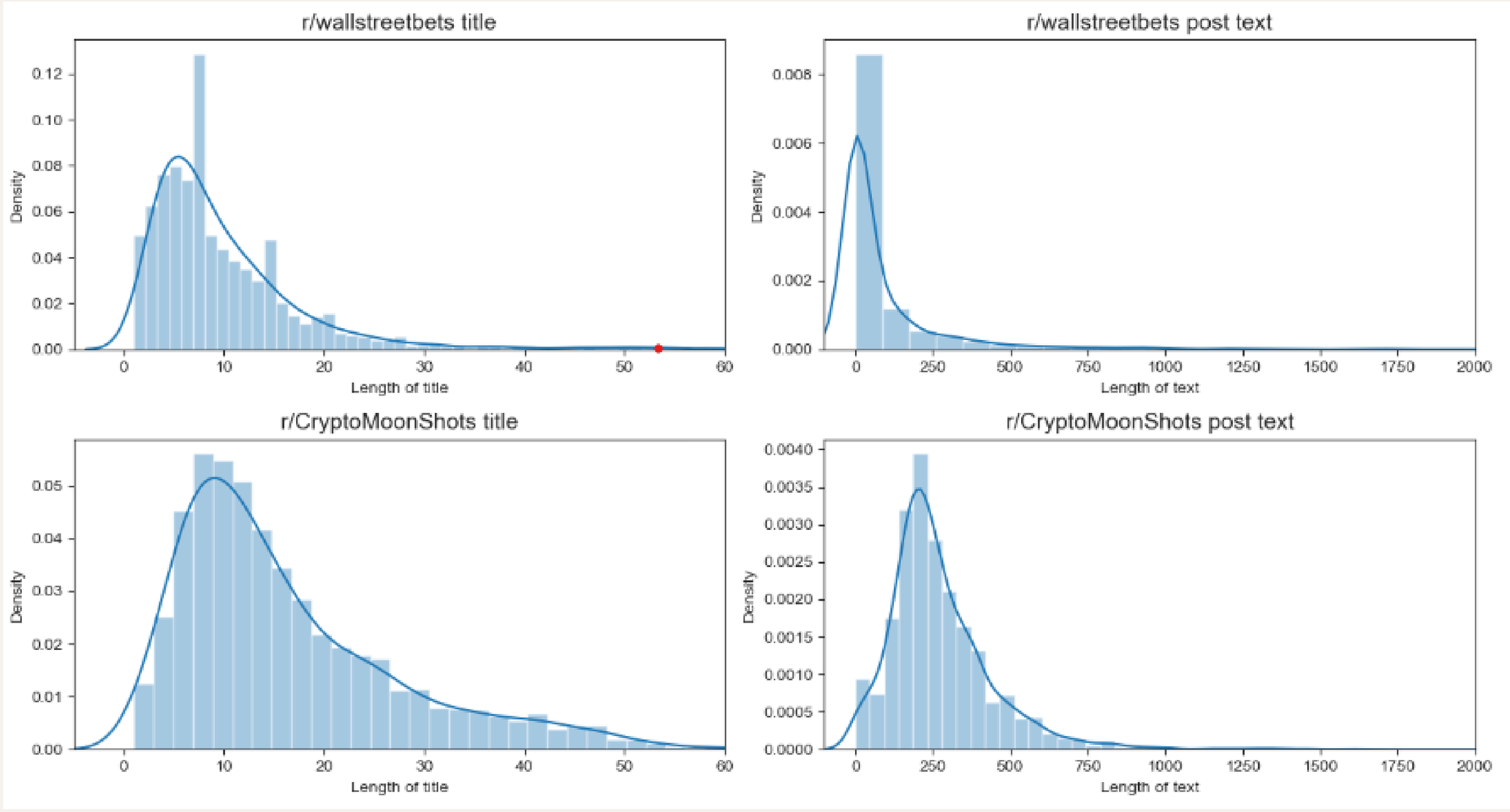
3.Train and compare classification models – multinomial naïve bayes, random forest trees and logistic regression - to predict whether a post is related to cryptocurrency or stocks.

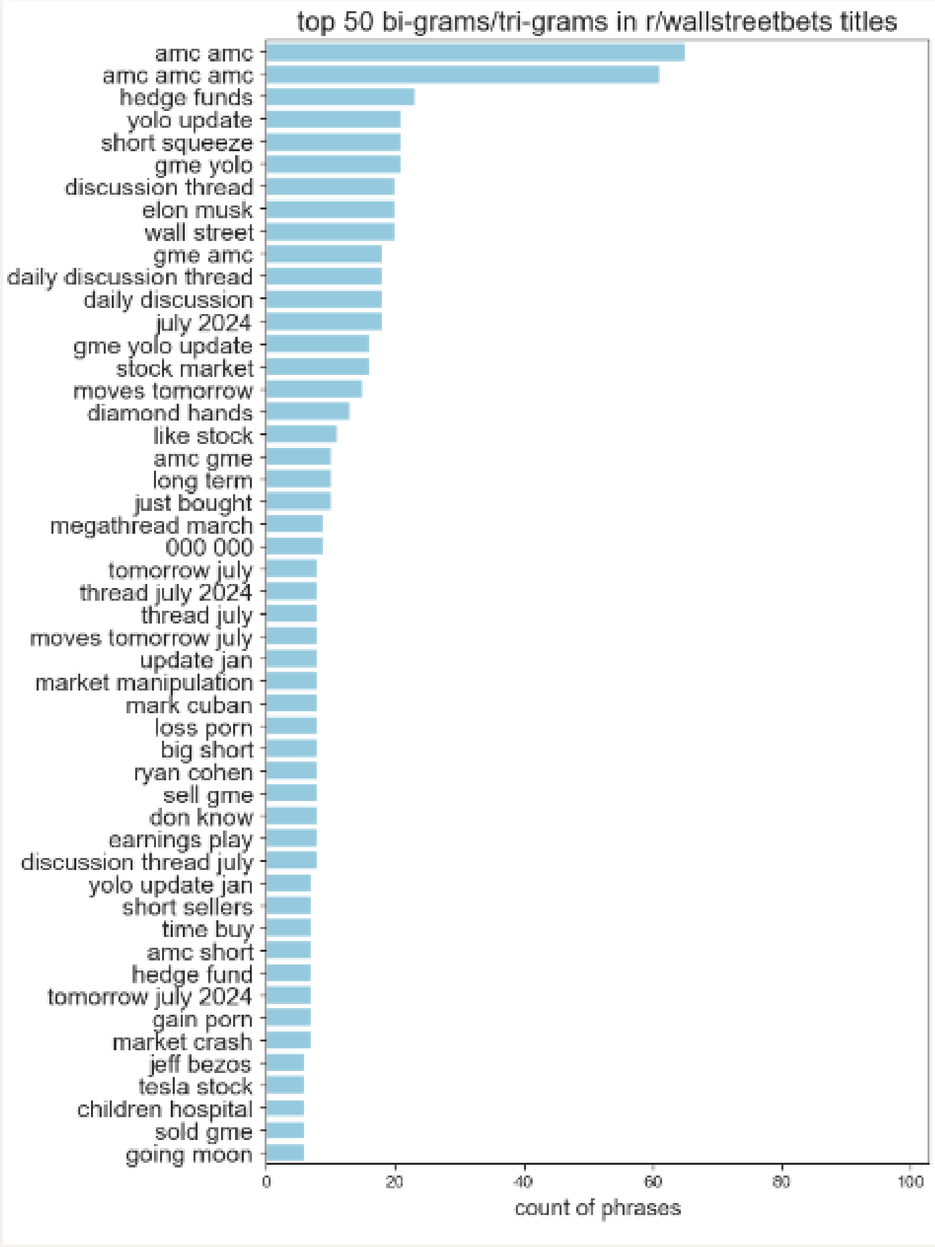
4.Evaluate the performance of the models using appropriate metrics such **as accuracy, precision, recall, and F1 score**.

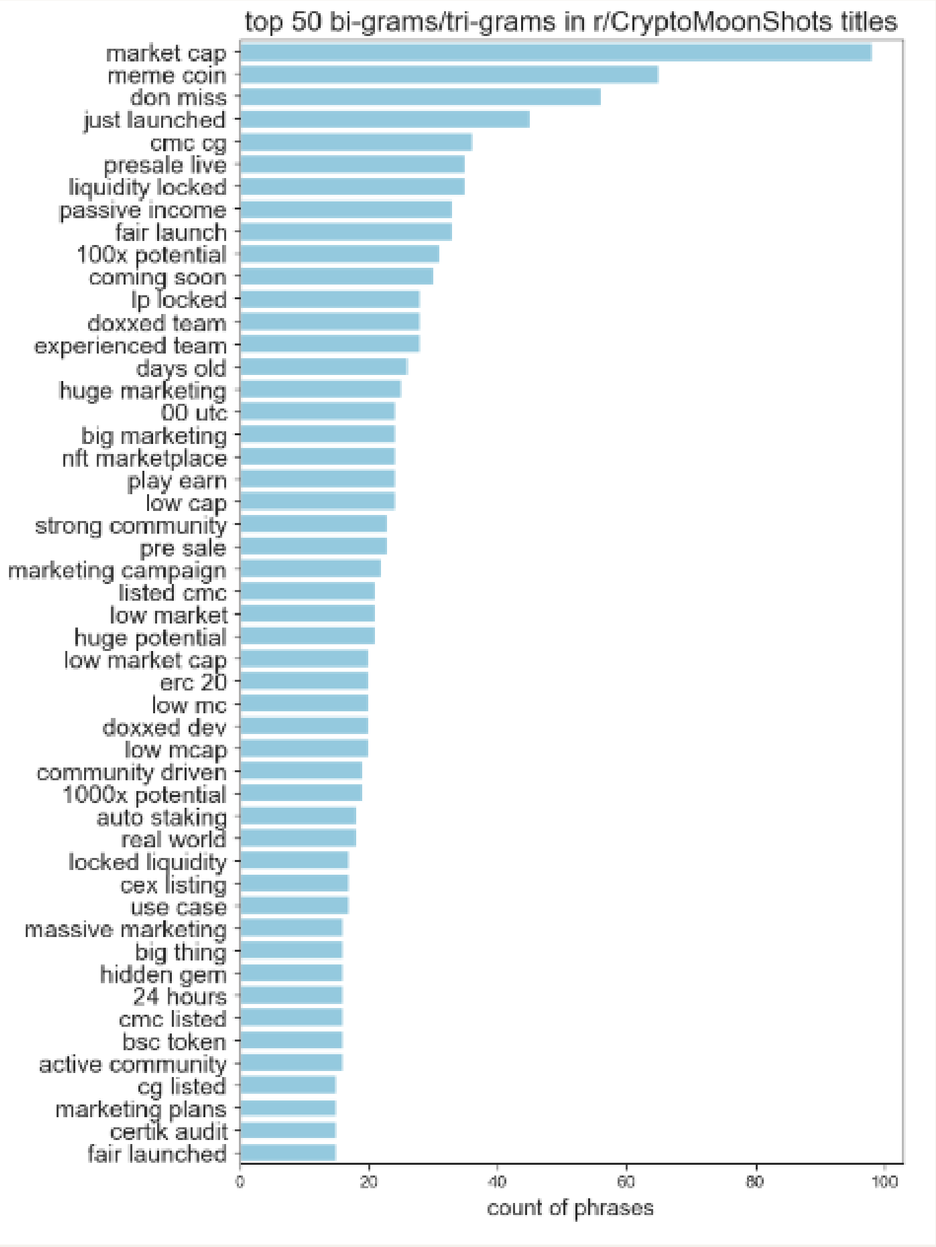


has_text	
False	51.292951
True	48.707049

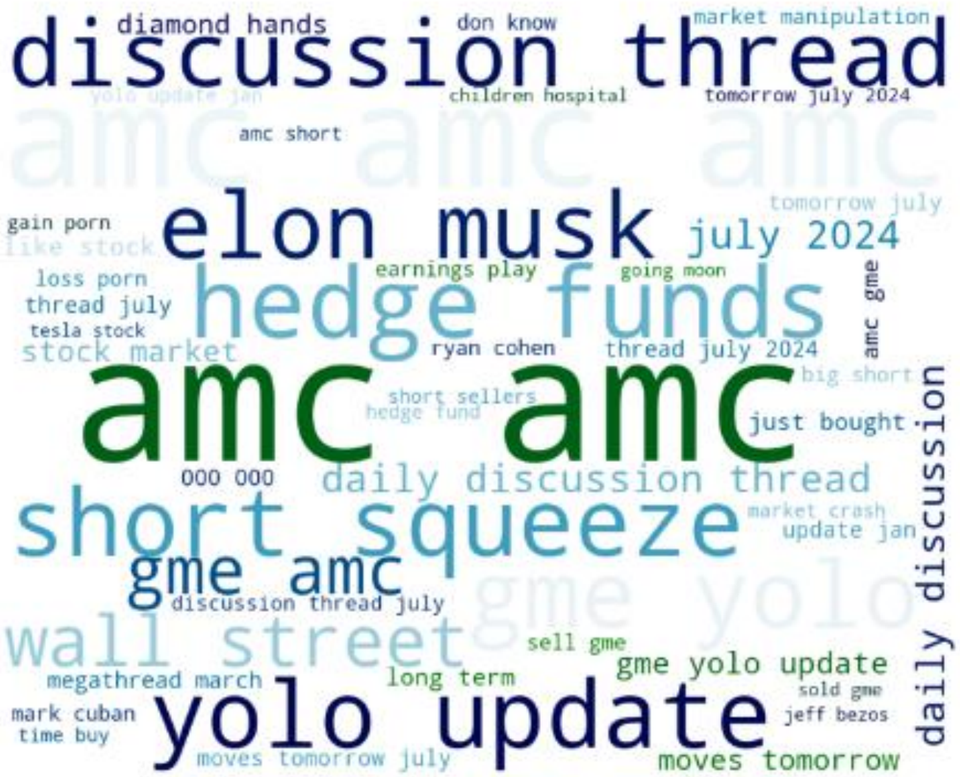
has_text	
True	99.375459
False	0.624541


r/wallstreetbets title
Density
Length of title
r/wallstreetbets post text
Density
Length of text
r/CryptoMoonShots title
Density
Length of title
r/CryptoMoonShots post text
Density
Length of text

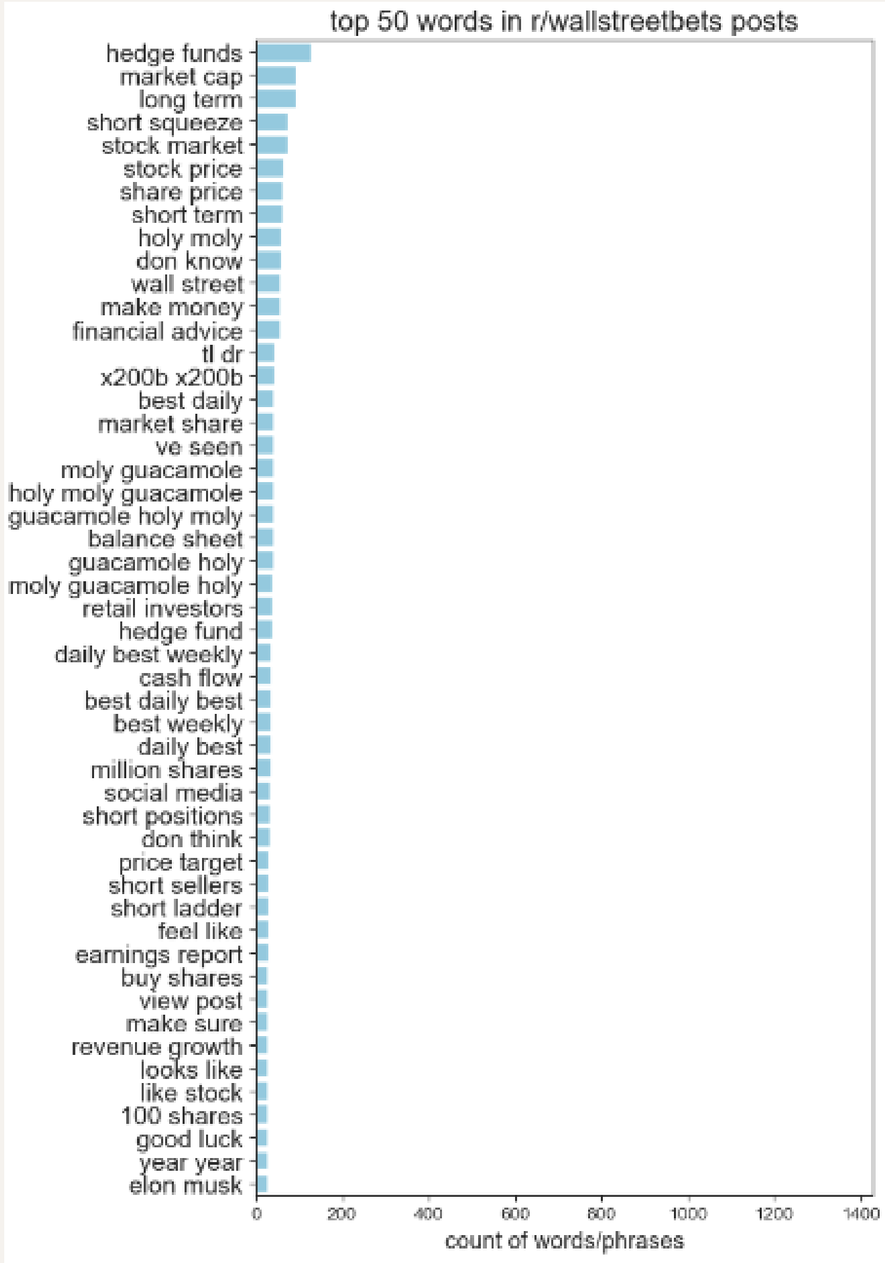

top 50 bi-grams/tri-grams in r/wallstreetbets titles
amc amc
amc amc amc
hedge funds
yolo update
short squeeze
gme yolo
discussion thread
elon musk
wall street
gme amc
daily discussion thread
daily discussion
july 2024
gme yolo update
stock market
moves tomorrow
diamond hands
like stock
amc gme
long term
just bought
megathread march
000 000
tomorrow july
thread july 2024
thread july
moves tomorrow july
update jan
market manipulation
mark cuban
loss porn
big short
ryan cohen
sell gme
don know
earnings play
discussion thread july
yolo update jan
short sellers
time buy
amc short
hedge fund
tomorrow july 2024
gain porn
market crash
jeff bezos
tesla stock
children hospital
sold gme
going moon
0
20
40
60
80
100
count of phrases

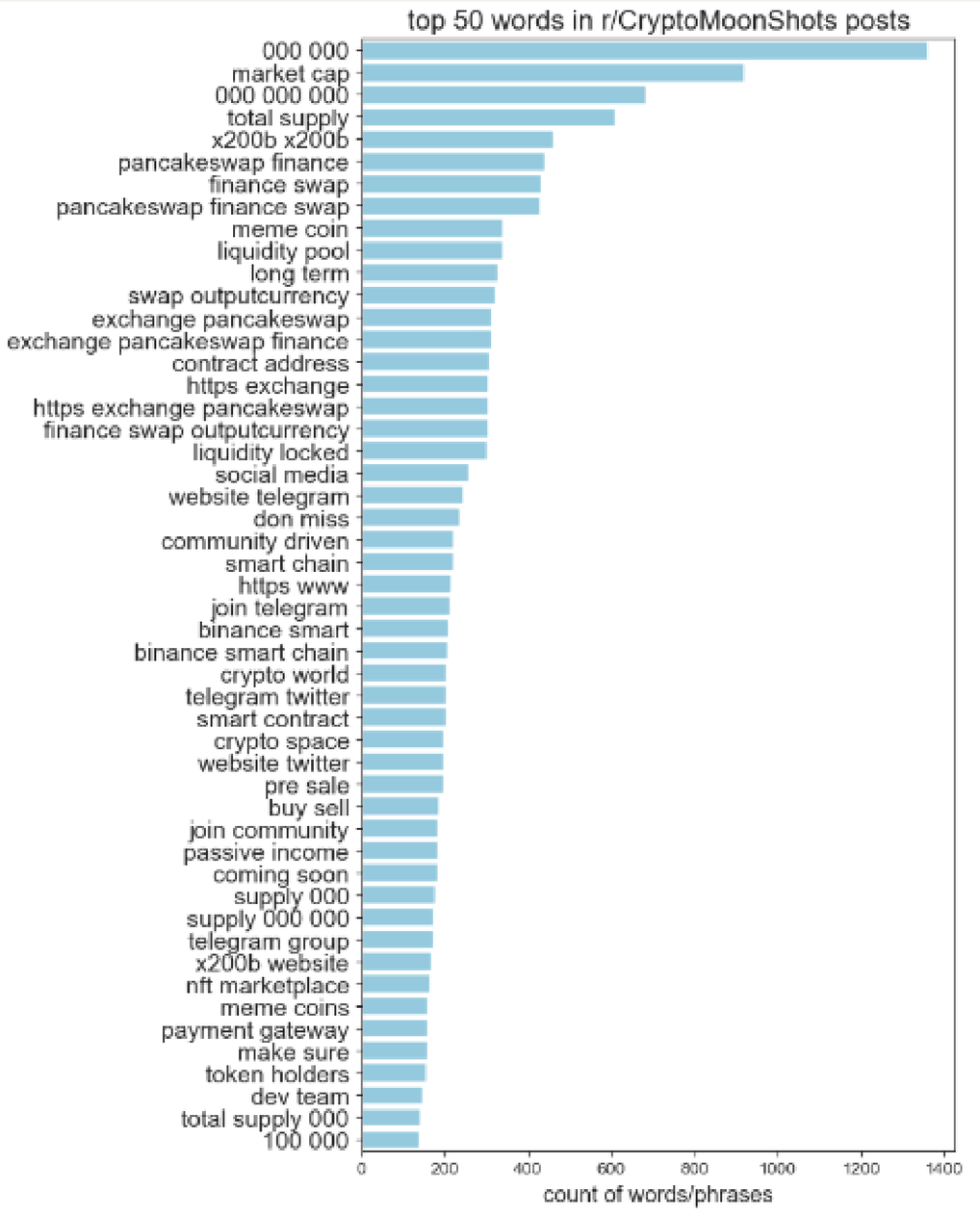

top 50 bi-grams/tri-grams in r/CryptoMoonShots titles
market cap
meme coin
don miss
just launched
cmc cg
presale live
liquidity locked
passive income
fair launch
100x potential
coming soon
lp locked
doxxed team
experienced team
days old
huge marketing
00 utc
big marketing
nft marketplace
play earn
low cap
strong community
pre sale
marketing campaign
listed cmc
low market
huge potential
low market cap
erc 20
low mc
doxxed dev
low mcap
community driven
1000x potential
auto staking
real world
locked liquidity
cex listing
use case
massive marketing
big thing
hidden gem
24 hours
cmc listed
bsc token
active community
cg listed
marketing plans
certik audit
fair launched
0
20
40
60
80
100
count of phrases

r/wallstreetbets

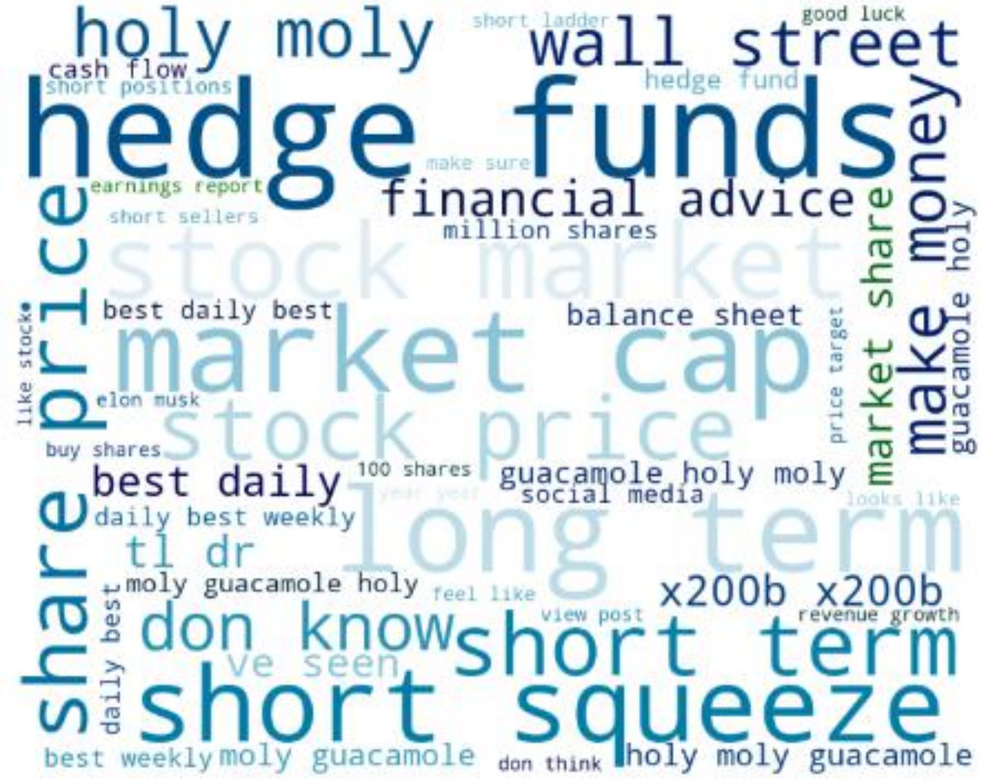


r/CryptoMoonShots

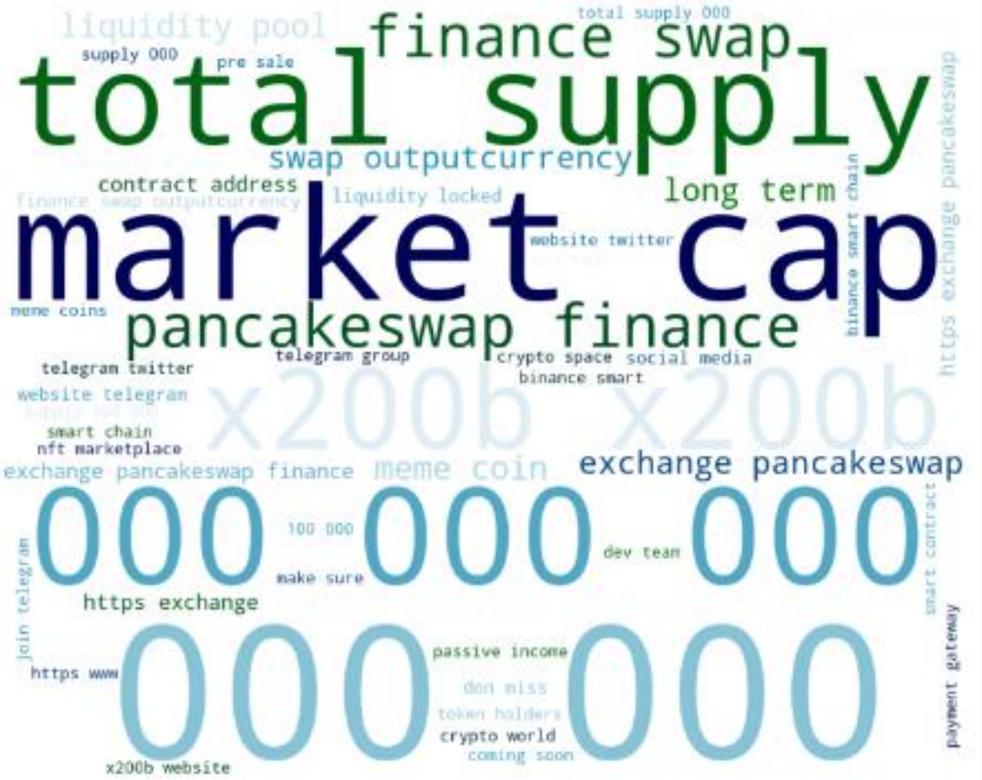

top 50 words in r/wallstreetbets posts
hedge funds
market cap
long term
short squeeze
stock market
stock price
share price
short term
holy moly
don know
wall street
make money
financial advice
tl dr
x200b x200b
best daily
market share
ve seen
moly guacamole
holy moly guacamole
guacamole holy moly
balance sheet
guacamole holy
moly guacamole holy
retail investors
hedge fund
daily best weekly
cash flow
best daily best
best weekly
daily best
million shares
social media
short positions
don think
price target
short sellers
short ladder
feel like
earnings report
buy shares
view post
make sure
revenue growth
looks like
like stock
100 shares
good luck
year year
elon musk
0
200
400
600
800
1000
1200
1400
count of words/phrases

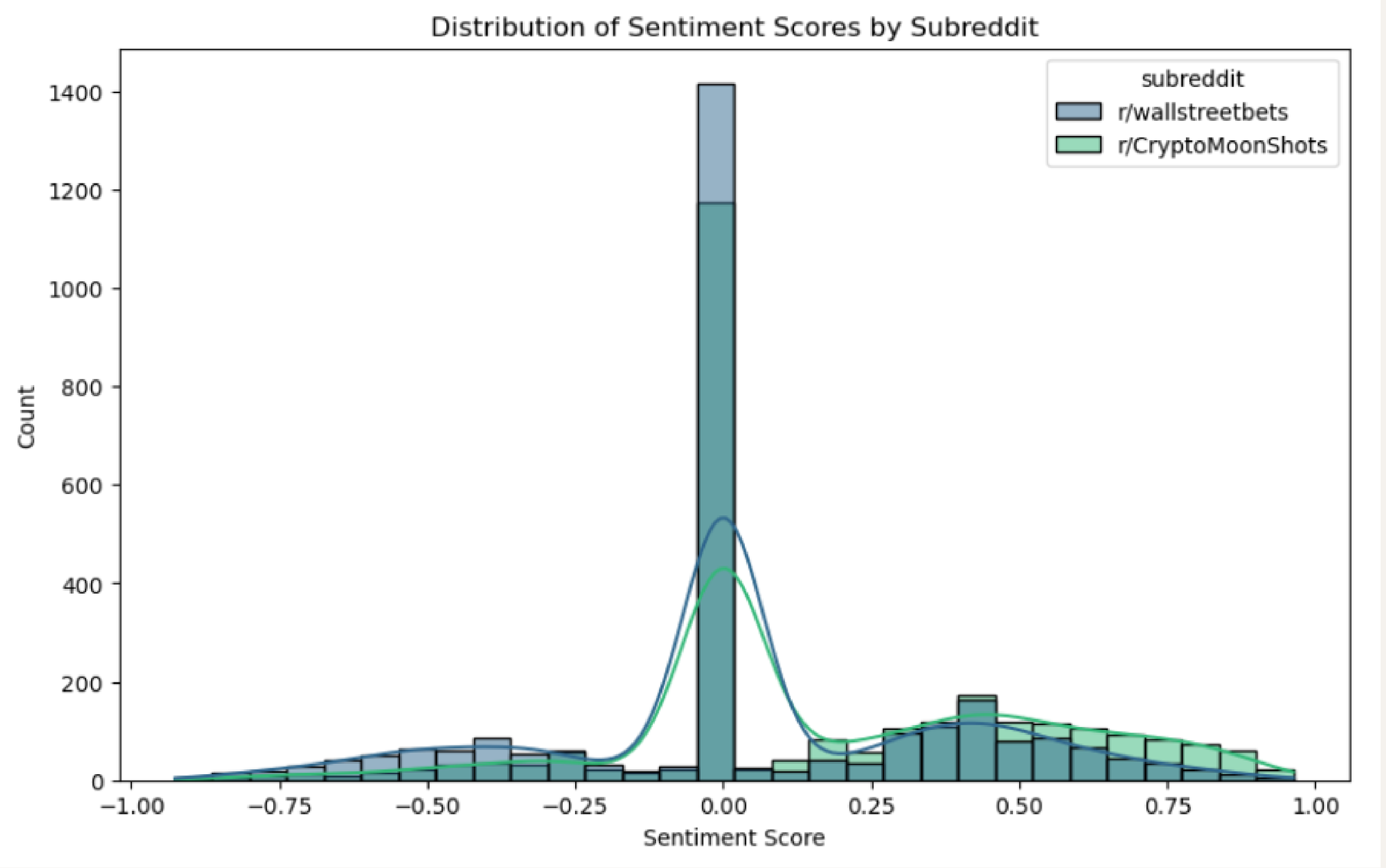

top 50 words in r/CryptoMoonShots posts
000 000
market cap
000 000 000
total supply
x200b x200b
pancakeswap finance
finance swap
pancakeswap finance swap
meme coin
liquidity pool
long term
swap outputcurrency
exchange pancakeswap
exchange pancakeswap finance
contract address
https exchange
https exchange pancakeswap
finance swap outputcurrency
liquidity locked
social media
website telegram
don miss
community driven
smart chain
https www
join telegram
binance smart
binance smart chain
crypto world
telegram twitter
smart contract
crypto space
website twitter
pre sale
buy sell
join community
passive income
coming soon
supply 000
supply 000 000
telegram group
x200b website
nft marketplace
meme coins
payment gateway
make sure
token holders
dev team
total supply 000
100 000
0
200
400
600
800
1000
1200
1400
count of words/phrases

r/wallstreetbets



r/CryptoMoonShots




Distribution of Sentiment Scores by Subreddit
subreddit
r/wallstreetbets
r/CryptoMoonShots
Count
1400
1200
1000
800
600
400
200
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Score

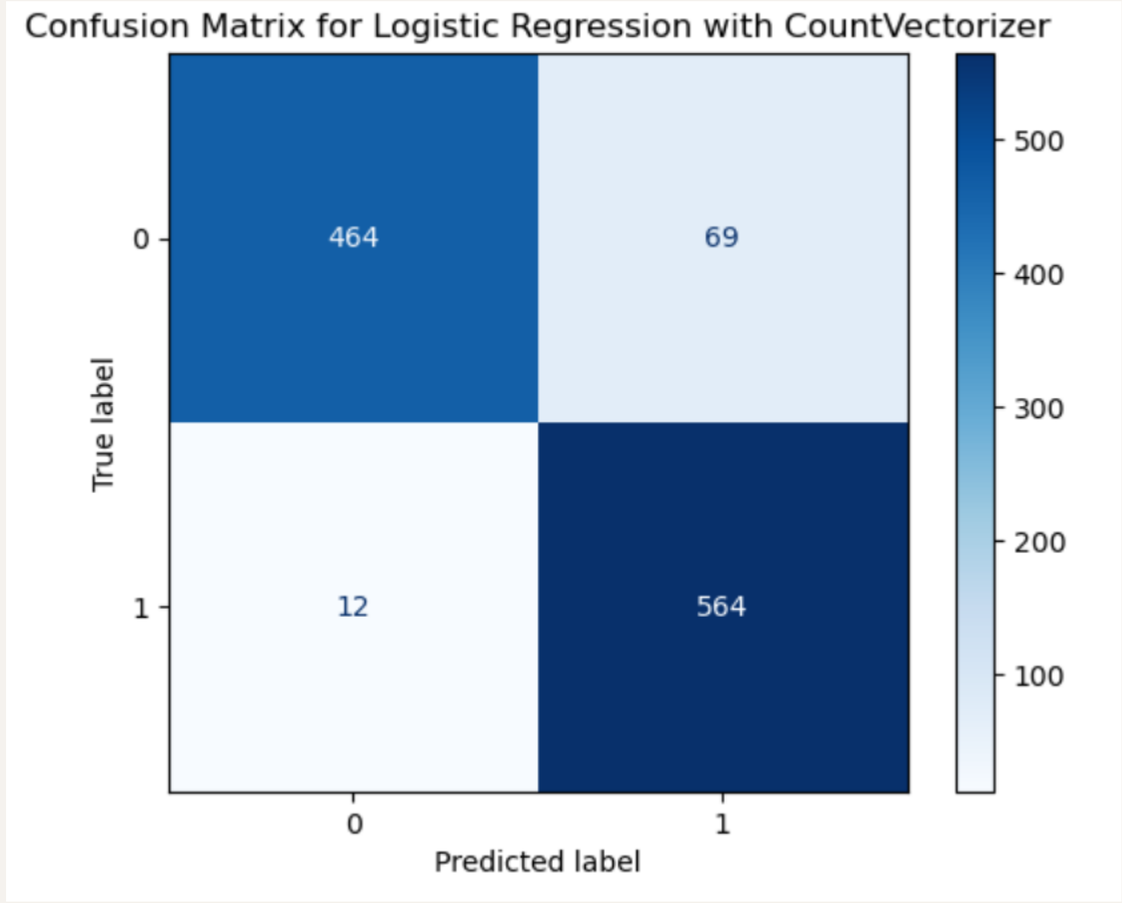
- After merging all dataframes into one combined csv file, mapping 0 to r/CryptoMoonShots and 1 to r/wallstreetbets, as well as tokenizing the title text into a processed feature for sentiment analysis, we can establish our baseline accuracy!
- **Baseline Accuracy – 0.519 or 51.9%**

Model 1: Logistic Regression with CountVectorizer

Training Accuracy	Testing Accuracy
0.928	0.927

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.97	0.87	0.92	533
1	0.89	0.98	0.93	576
accuracy			0.93	1109
macro avg	0.93	0.92	0.93	1109
weighted avg	0.93	0.93	0.93	1109

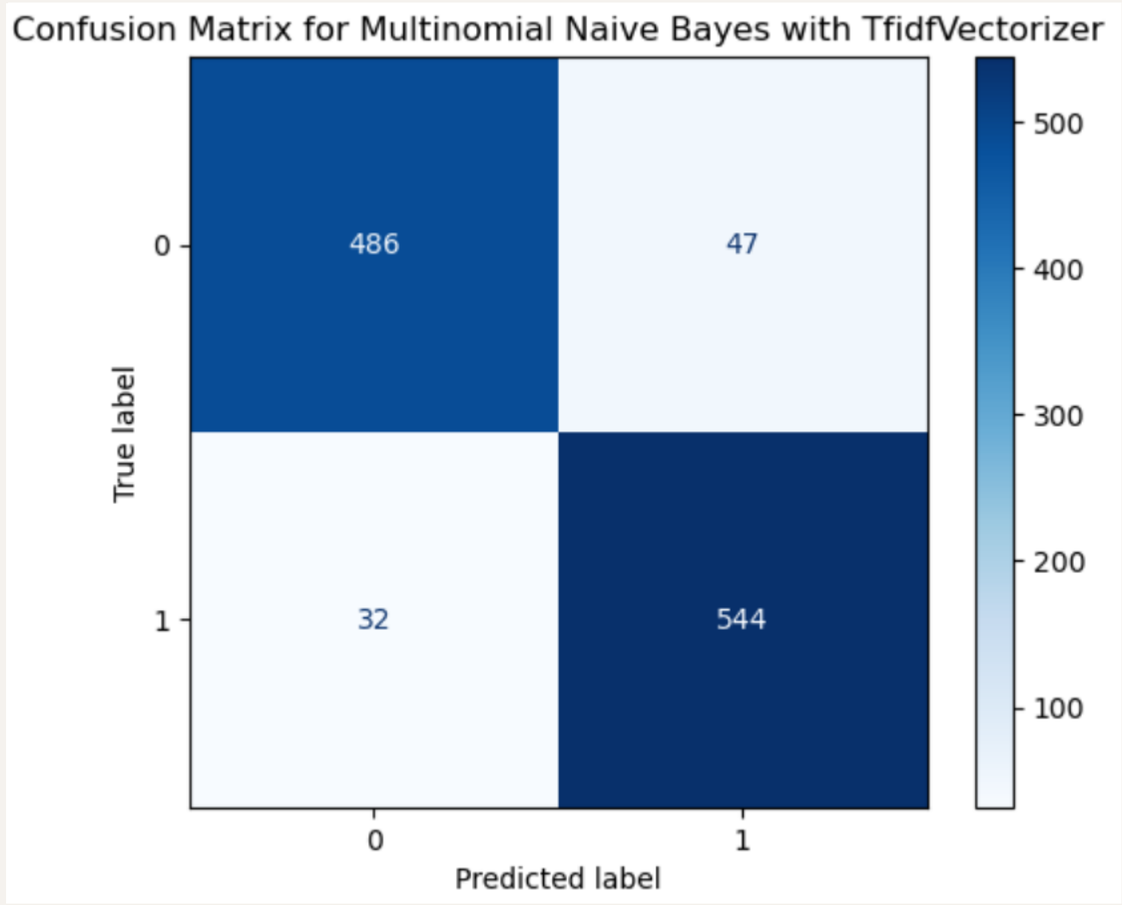


Model 2: Multinomial Naïve Bayes with TfidfVectorizer

Training Accuracy	Testing Accuracy
0.928	0.929

Classification Report for Multinomial Naive Bayes:

	precision	recall	f1-score	support
0	0.94	0.91	0.92	533
1	0.92	0.94	0.93	576
accuracy			0.93	1109
macro avg	0.93	0.93	0.93	1109
weighted avg	0.93	0.93	0.93	1109

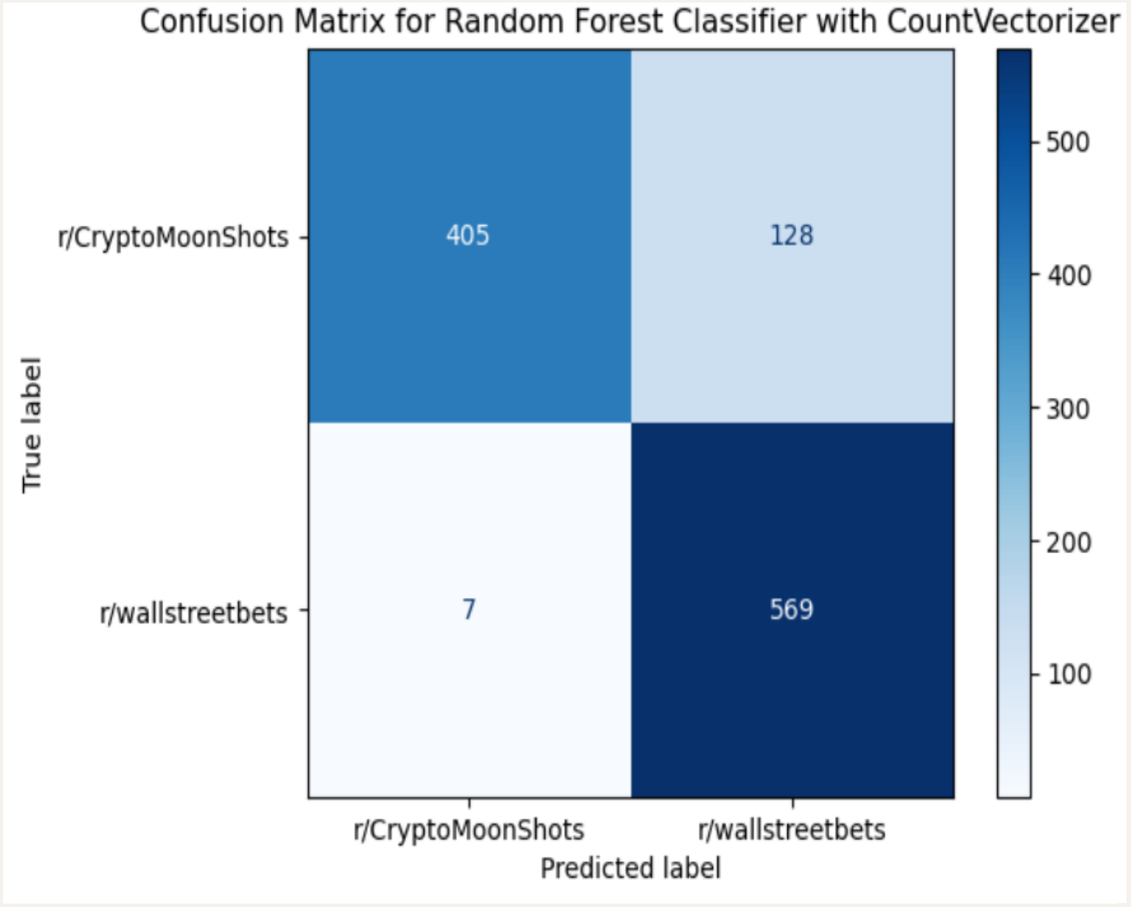


Model 3: Random Forest Tree Classifier with CountVectorizer

Training Accuracy	Testing Accuracy
0.891	0.878

Classification Report for Random Forest Classifier:

	precision	recall	f1-score	support
0	0.98	0.76	0.86	533
1	0.82	0.99	0.89	576
accuracy			0.88	1109
macro avg	0.90	0.87	0.88	1109
weighted avg	0.90	0.88	0.88	1109

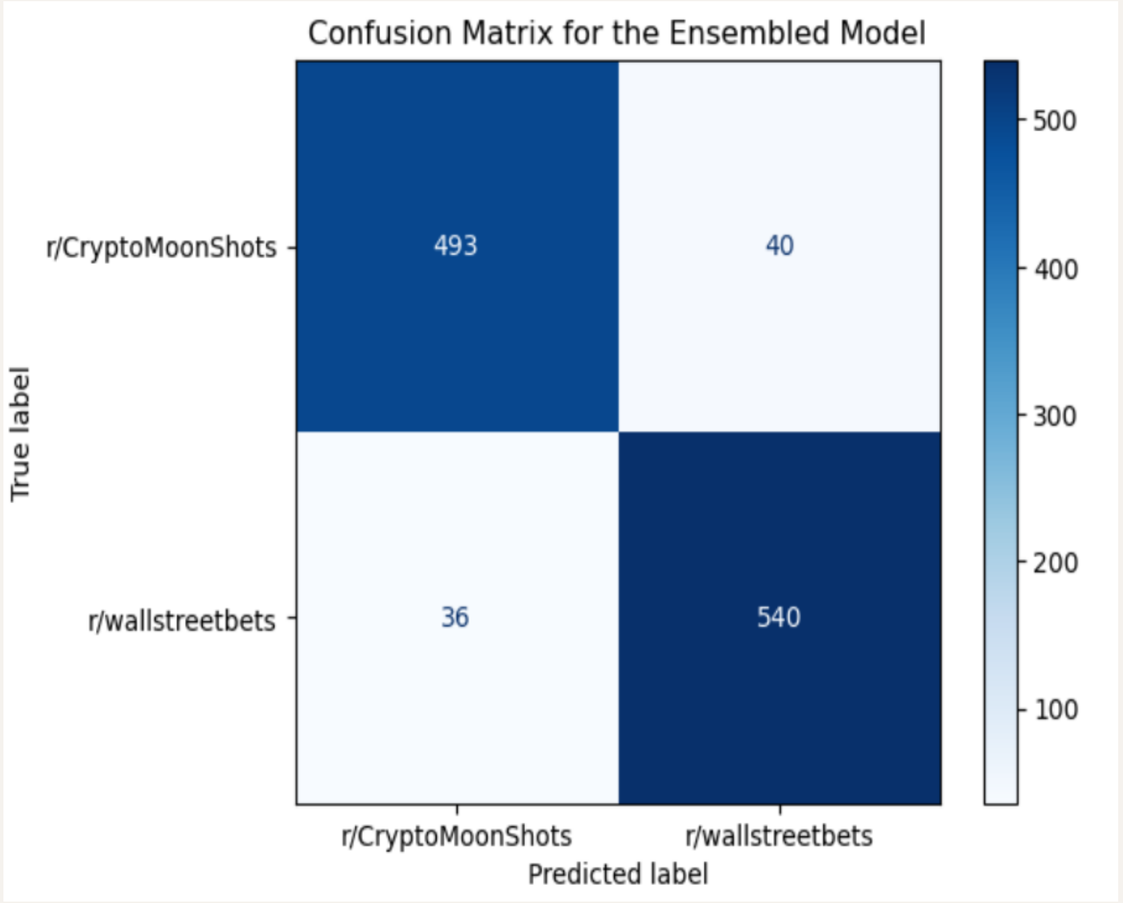


Model 4: Ensemble Model (Logistic Regression & Multinomial Naïve Bayes)

Testing Accuracy
0.931

Classification Report for Ensembled Model:

	precision	recall	f1-score	support
0	0.93	0.92	0.93	533
1	0.93	0.94	0.93	576
accuracy			0.93	1109
macro avg	0.93	0.93	0.93	1109
weighted avg	0.93	0.93	0.93	1109



Conclusions/Insights

- **Precision, Recall, F1-score, and Accuracy:** The ensembled model has similar performance to the first two individual models, with **high precision of 0.93**, **recall of around 0.93**, **F1-score of 0.93,** for both subreddits and an overall **accuracy score of 0.93**.

- This will give more of a balance with the false negatives and positives; with **36 false negatives vs 40 false positives**! Let's **choose this ensemble model** for production!

Thanks!