

Project 3: Subreddit Classification

John Gilheany



Table of Contents



01

Problem Statement

02

Data Overview

03

Model Comparisons

04

Final Recommendation



01

Problem Statement



Problem Statement

- Our consulting firm, General Assembly Solutions (GAS), has been hired by Robinhood Markets, Inc. to advise on the marketing of its new cryptocurrency offering.
- Robinhood would like our help in better understanding the content and language differences between public markets and crypto subreddits to ultimately build a classification model that can predict which subreddit a post belongs to.



02


Data Overview



WallStreetBets (WSB)


Posted by u/LCDRtomdodge 14 hours ago

210 **Calls on \$TSLA?** Meme

 **r/EnoughMuskSpam**
u/LifeResolution7 · 12h

Tesla on course for \$40 billion value wipeout as shares plunge following Elon Musk's downbeat Cybertruck outlook


Who Needs Profits?



markets.businessinsider.com Open

40 billion (Tesla) + 44 billion (twitter) = 84 Billion all, basically, due to ONE man thinking he is ALWAYS the smartest person in the ROOM.

83 Comments Share Save ...

 **PINNED BY MODERATORS**

35 Posted by u/OPINION_IS_UNPOPULAR AutoModerator's Father 4 hours ago

Daily Discussion Thread for October 20, 2023 Daily Discussion

1.6k Comments Share Save ...

Posted by u/rylar Supports a arfield Ethnostate 14 hours ago

83 **Most Anticipated Earnings Releases for the week beginning October 23rd, 2023** Earnings Thread

59 Comments Share Save ...



Posted by u/LiquidatedAF Liquidated Jensen Huang 17 hours ago




224 **ENPH gains** Gain



Position	Days Left	YTD Mkt Gain	Market Value	Cr 1/3 Value
+30 Oct23 \$127 Put X 10/20, IN	1	\$27,510.00	33,600.00	6,090.00



181 Comments Share Save ...




CryptoMoonShots (CMS)



  PINNED BY MODERATORS




7.5k Posted by u/LucidDreamState 3 years ago   


 **Introducing the official CryptoMoonShots premium Discord!** 



 2 Comments  Share  Save ...

 2.5k Posted by u/LucidDreamState 5 years ago   

 **Read the sidebar before making a new submission** 

 4 Comments  Share  Save ...

 23 Posted by u/rs56j5jr45sjr4 18 hours ago

 **The Revolutionary Arsenal 2.0: Elevating Your Gaming Experience | Founded 2020** 


Gamers and blockchain enthusiasts, hold onto your seats because a seismic shift is coming to the gaming world. Arsenal 2.0 is poised to redefine your gaming experience when it launches this 3rd of November. From weapon NFTs to tantalizing new game modes, here's why you should be buzzing with excitement.



What's Hot?

Weapon NFTs: An Arsenal Like No Other

Bland, generic guns are a thing of the past. Arsenal 2.0 introduces weapon NFTs, allowing you not only to own distinct weapons but also to trade or sell them on the blockchain. You're not just collecting guns; you're amassing valuable, one-of-a-kind assets.

Customize NFTs: Your Gun, Your Rules

 604 Posted by u/hugikea 21 hours ago


 **Wednesday Inu | Elon Musk just tweeted | Huge potential | Buy Now** 


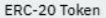
Welcome to Wednesday Inu

Elon Musk made a Tweet and a hidden advertisement for the Wednesday token. Elon Musk bought a Wednesday token

Tweet - <https://x.com/elonmusk/status/1714519142617805018?s=20>

The Wednesday token has increased in price by 60% and is rising even higher. Perhaps the Wednesday token will overcome ATH in the coming days

 1 Posted by u/Siddy676 4 days ago

 **The Top 5 Cryptocurrencies With Rock-Solid Fundamentals For 2023 And 2024** 

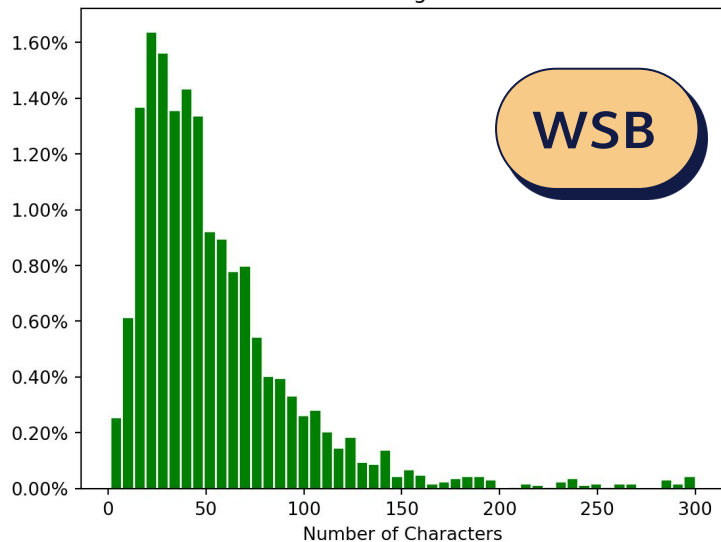
In the dynamic landscape of cryptocurrency, being informed is crucial. With thorough research, we have compiled a list of the top five cryptocurrencies set to make an impact in 2023 and 2024. Our selections are based on solid fundamentals, active community engagement, and promising growth prospects.

- Advertisement -

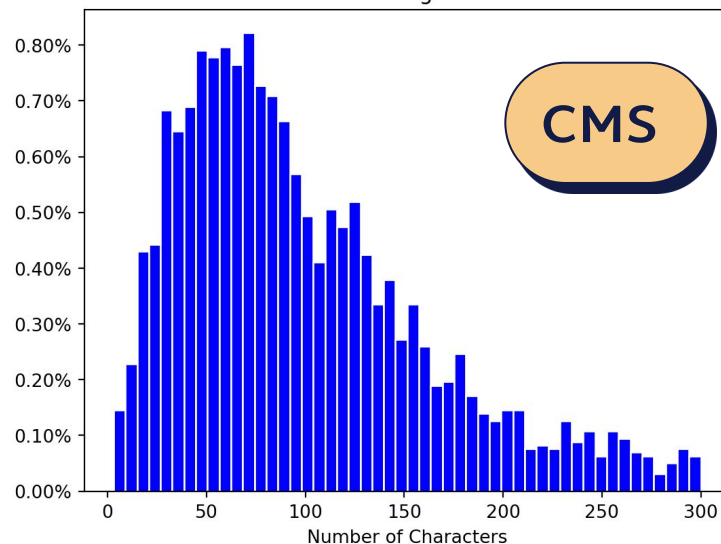
SafePal (SFP)

Title Length by Subreddit

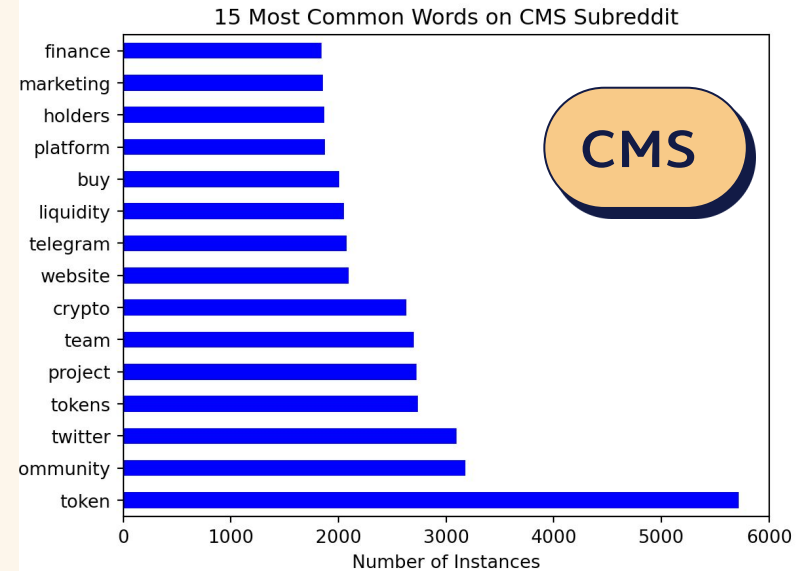
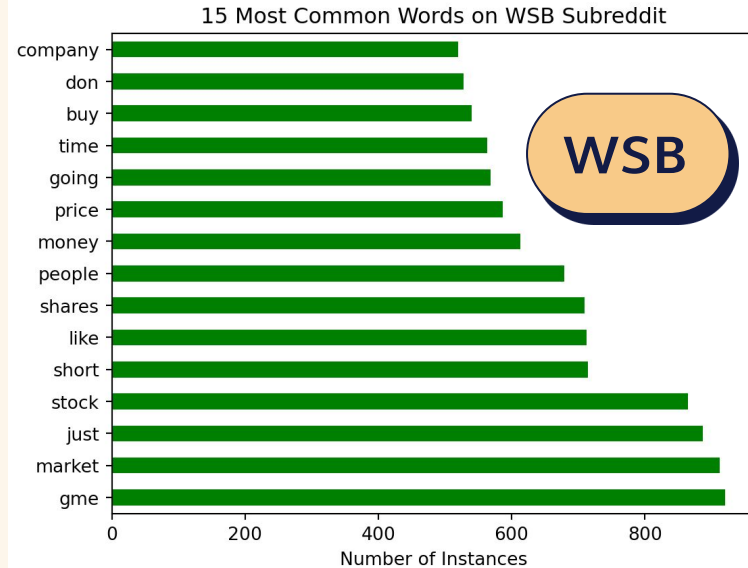
Distribution of Title Length in WSB Subreddit



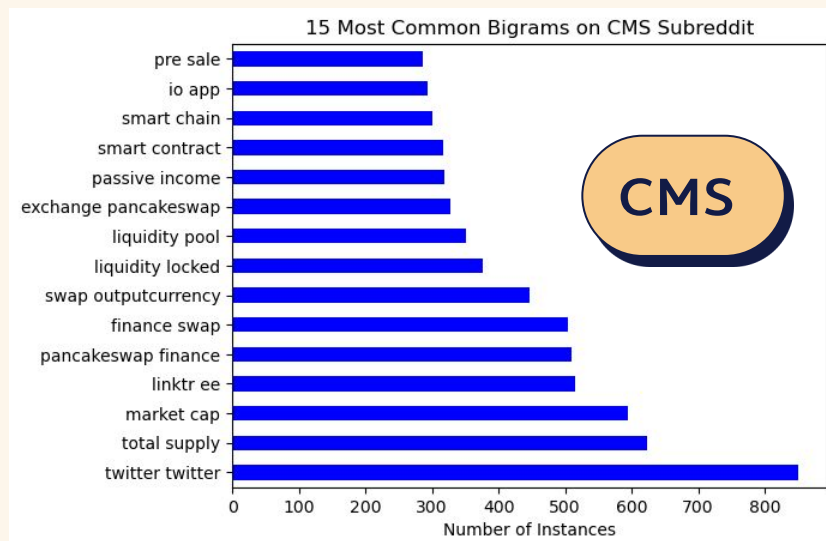
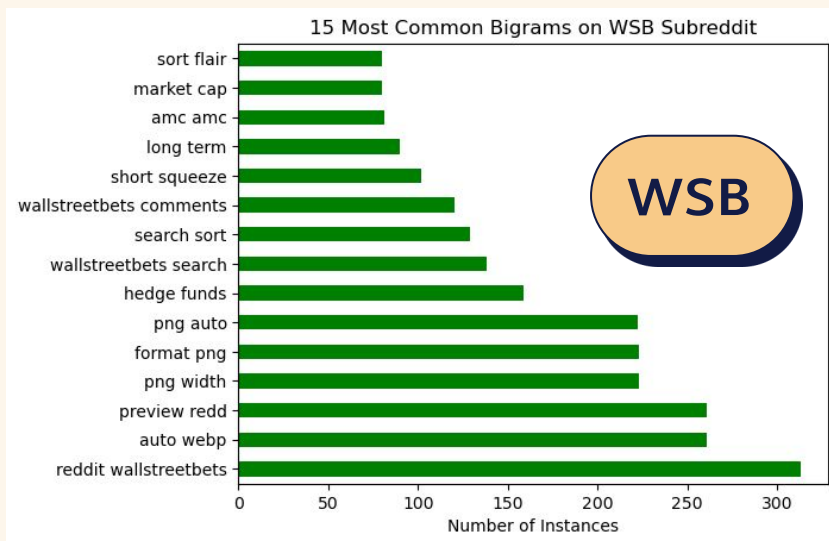
Distribution of Title Length in CMS Subreddit



15 Most Common Words by Subreddit



15 Most Common Bigrams by Subreddit



03

Model Comparisons



Model #1: Multinomial Naive Bayes (Count Vectorizer) - Accuracy: 97.2%

Best Parameters:

'cvec__stop_words': None,
'cvec__ngram_range': (1, 1),
'cvec__min_df': 2,
'cvec__max_features': None,
'cvec__max_df': 0.95,
'cvec__binary': True

Train Score: 0.973944

Test Score: 0.971927

		Predicted Classification	
		WSB	CMS
Actual Classification	WSB	623	24
	CMS	13	658

Recall	98.1%
Accuracy	97.2%
Precision	96.3%
F1 Score	0.97

Model #2: Multinomial Naive Bayes (Tfidf Vectorizer) - Accuracy: 96.9%

Best Parameters:

'tvec__stop_words': None,
'tvec__ngram_range': (1, 1),
'tvec__min_df': 3,
'tvec__max_features': 7000,
'tvec__max_df': 0.95,
'tvec__binary': True

Train Score: 0.969896

Test Score: 0.968892

**Actual
Classification**

WSB
CMS

Predicted Classification

WSB	CMS
618	29
12	659

Recall	98.2%
Accuracy	96.9%
Precision	95.5%
F1 Score	0.97

Model #3: K-Nearest Neighbors (KNN) - Accuracy: 89.0%

Best Parameters:

'knn__weights': 'uniform',
'knn__n_neighbors': 5,
'cvec__stop_words': None,
'cvec__ngram_range': (1, 1),
'cvec__min_df': 1,
'cvec__max_features': 2000,
'cvec__max_df': 0.9,
'cvec__binary': False

Train Score: 0.928156

Test Score: 0.889985

**Actual
Classification**

WSB
CMS

Predicted Classification

WSB CMS

626	21
124	547

Recall 81.5%

Accuracy 89.0%

Precision 96.8%

F1 Score 0.88

Model #4: Logistic Regression (LogReg) - Accuracy: 98.6%

Best Parameters:

'logreg__penalty': None,
'cvec__stop_words': 'english',
'cvec__ngram_range': (1, 1),
'cvec__min_df': 1,
'cvec__max_features': None,
'cvec__max_df': 0.95,
'cvec__binary': False

Train Score: 1.0

Test Score: 0.986343

**Actual
Classification**

WSB
CMS

Predicted Classification

WSB CMS

643	4
14	657

Recall 97.9%

Accuracy 98.6%

Precision 99.4%

F1 Score 0.99

Model #5: Random Forest (RF) - Accuracy: 97.8%

Best Parameters:

'rf__n_estimators': 100,
'rf__max_features': 'sqrt',
'rf__criterion': 'log_loss',
'cvec__stop_words': 'english',
'cvec__ngram_range': (1, 2),
'cvec__min_df': 1,
'cvec__max_features': 2000,
'cvec__max_df': 0.9,
'cvec__binary': False

Train Score: 0.999241

Test Score: 0.977997

**Actual
Classification**

WSB
CMS

Predicted Classification

WSB	CMS
640	7
22	649

Recall 96.7%

Accuracy 97.8%

Precision 98.9%

F1 Score 0.98

Model #6: Extra Trees Classifier (ET) - Score: Accuracy: 98.0%

Best Parameters:

```
'et__n_estimators': 500,  
'et__min_samples_leaf': 2,  
'et__criterion': 'log_loss',  
'cvec__stop_words': 'english',  
'cvec__ngram_range': (1, 1),  
'cvec__min_df': 2,  
'cvec__max_features': 2000,  
'cvec__max_df': 0.95,  
'cvec__binary': True
```

Train Score: 0.994435

Test Score: 0.979514

**Actual
Classification**

WSB
CMS

Predicted Classification

WSB	CMS
644	3
24	647

Recall 96.4%

Accuracy 98.0%

Precision 99.5%

F1 Score 0.98

04

Final Recommendation





Best Model: Logistic Regression

Model	Accuracy	Recall	Precision
Logistic Regression (LogReg)	98.6%	97.9%	99.4%
Extra Trees Classifier (ET)	98.0%	96.4%	99.5%
Random Forest (RF)	97.8%	96.7%	98.9%
Multinomial Naive Bayes (Count Vectorizer)	97.2%	98.1%	96.3%
Multinomial Naive Bayes (Tfid Vectorizer)	96.9%	98.2%	95.5%
K-Nearest Neighbors (KNN)	89.0%	81.5%	96.8%



Thanks!

Any questions?



CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

