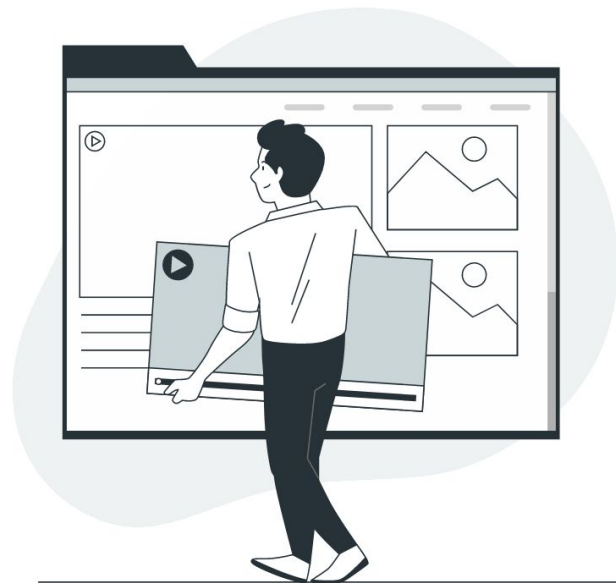


Classifying Posts from Two Subreddits: Bitcoin and Ethereum

DSI-23: Project 3

Deepankar Sharma
Raymond Onn
Ray Tan
Ash Ang



Contents

01 Introduction

Deepankar Sharma

02 Methods & Analysis

Raymond Onn



03 Vectorizing & Modelling

Ray Tan

04 Findings & Suggestions

Ash Ang

CryptoTrade23



Bitcoin (BTC)
logo



Ethereum (ETH)
logo



XRP (XRP) logo



Bitcoin Cash
(BCH) logo



Litecoin (LTC)
logo



EOS (EOS) logo



Binance Coin
(BNB) logo



Tether (USDT)
logo



Stellar (XLM)
logo



Cardano (ADA)
logo



TRON (TRX)
logo



Monero (XMR)
logo



Dash (DASH)
logo



Bitcoin SV (BSV)
logo



IOTA (MIOTA)
logo



Tezos (XTZ)
logo



Ethereum
Classic (ETC)
logo



NEO (NEO) logo



Ontology (ONT)
logo



Maker (MKR)
logo



NEM (XEM) logo



Basic Attention
Token (BAT)
logo



Zcash (ZEC)
logo



Bitcoin Gold
(BTG) logo



VeChain (VET)
logo



USD Coin
(USDC) logo



Dogecoin
(DOGE) logo



Decred (DCR)
logo



Waves (WAVES)
logo



OmiseGO (OMG)
logo

Background

CryptoTrade23 is a fintech startup specialising in cryptocurrency investments.

- **Customer Service Team** has been receiving an overwhelming number of enquiries about cryptocurrencies.
- Head of Customer Service has engaged the **Data Team** to automate responses to simple enquiries in the face of increasing workload and resource constraints.
- Enable **Customer Service Team** to focus on more complex enquiries.

Problem Statement

We aim to build a chatbot for the company's website as the first checkpoint for users enquiring about cryptocurrencies through the application of Natural Language Processing (NLP) and Machine Learning (ML) Classifiers.

The classifier in the chatbot will be trained to **respond to these enquiries based on keywords** in the users' inputs.



Objectives / Rationales

In this project, our team will be using data from two subreddits to achieve the following:

1. **Analyse and understand the text data** in each of the subreddits using Natural Language Processing (NLP).
2. Use Machine Learning (ML) classifiers to **identify the subreddit** a submission is likely to originate from.
3. **Evaluate the ML classifiers** against our baseline model using accuracy and ROC AUC as the metrics
4. Propose a suitable optimal ML classifier that could be used to **develop a minimum viable product (MVP) for the chatbot** and make other recommendations.

Subreddits Chosen / Quantity Of Data

Bitcoin

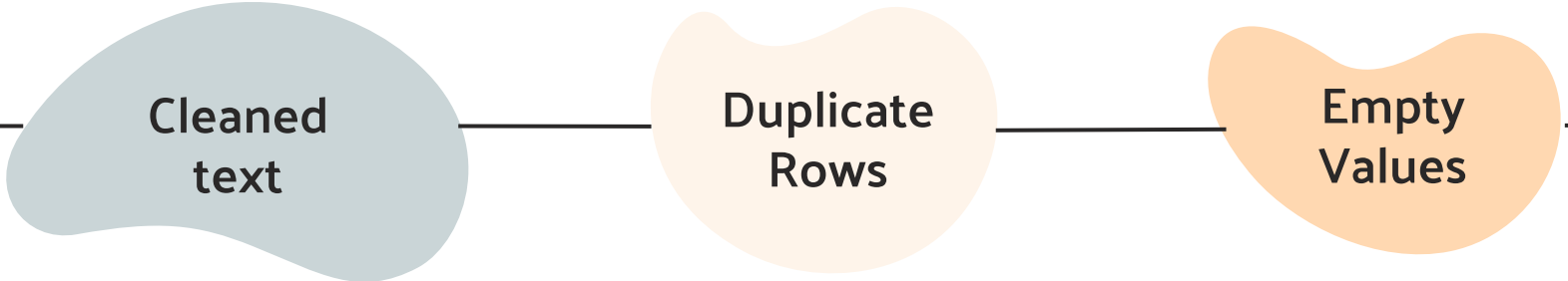
- 5,000 most recent posts (as of 22nd July 2021)
- <https://www.reddit.com/r/Bitcoin/>

Ethereum

- 10,000 most recent posts (as of 22nd July 2021)
- <https://www.reddit.com/r/ethereum>



Data Cleaning



Cleaned text

Removed html tags,
url links, punctuations
and words with 2 or
fewer letters

Duplicate Rows

Removed duplicate
values (about 4% of
sample)

Empty Values

Empty posts were
also removed

After cleaning, we were left with 2700 rows,
with **58%** of rows coming from r/Bitcoin
and **42%** of rows coming from r/Ethereum

Data Dictionary

	subreddit	post_title	post_content
0	Bitcoin	what moves crypto market apart from the speculators	would like know there anything that moves crypto market apart from the speculators ...
1	Bitcoin	help starting crypto business	guys interested starting crypto business app that would take small amounts money f...
2	Bitcoin	did jack dorsey confirm deflect taking btc for advertising	might just being stupid but watching the conference there the moment elon asks dorsey about a...

subreddit

Post_title (title)


4


 **r/Bitcoin** Posted by u/hawk-fe 7 days ago

what moves crypto market apart from the speculators

I would like to know if there is anything that moves crypto market apart from the speculators, or are cryptocurrencies and their prices absolutely speculative?

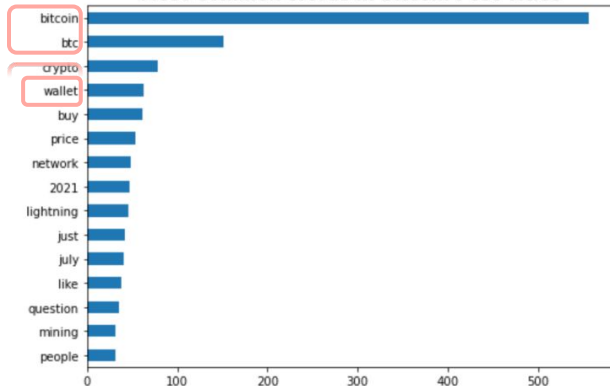
 **7 Comments**  **Award**  **Share** ...

83% Upvoted

Post_content (selftext)

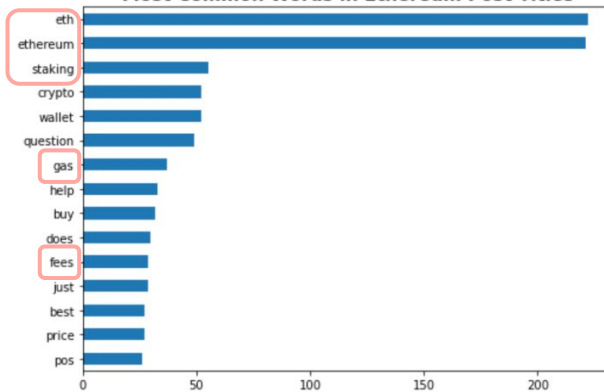
Top Words: Post Title

Most Common Words in Bitcoin Post Titles

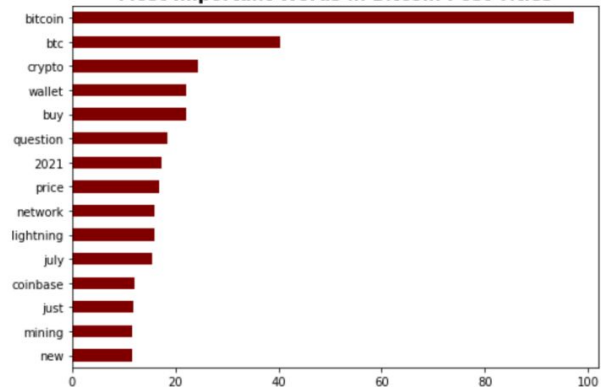


MOST
COMMON

Most Common Words in Ethereum Post Titles

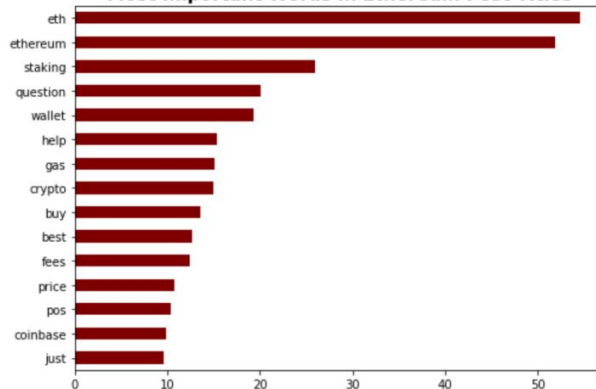


Most Important Words in Bitcoin Post Titles



MOST
IMPORTANT

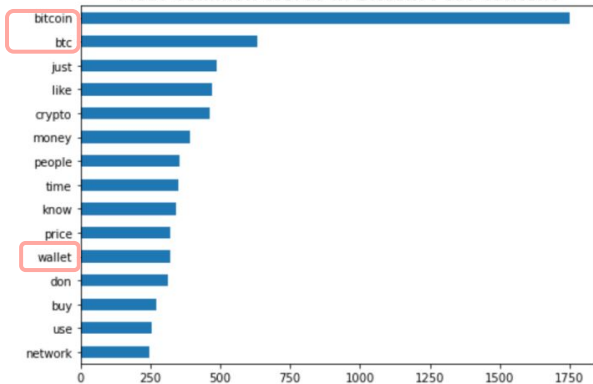
Most Important Words in Ethereum Post Titles



Note: Word importance was
determined based on calculated
TF-IDF

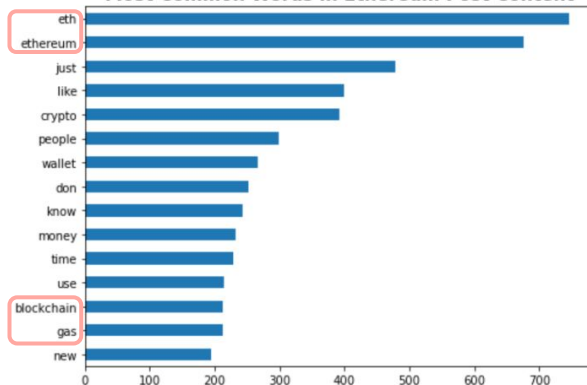
Top Words: Post Content

Most Common Words in Bitcoin Post Content

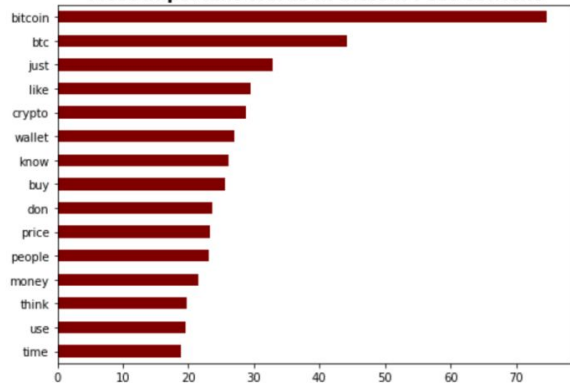


MOST
COMMON

Most Common Words in Ethereum Post Content

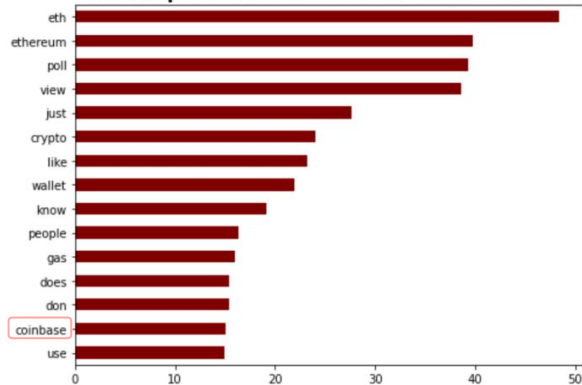


Most Important Words in Bitcoin Post Content



MOST
IMPORTANT

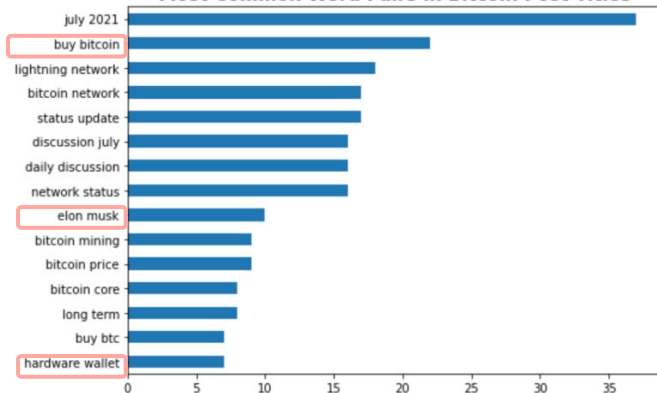
Most Important Words in Ethereum Post Content



Note: Word importance was determined based on calculated TF-IDF

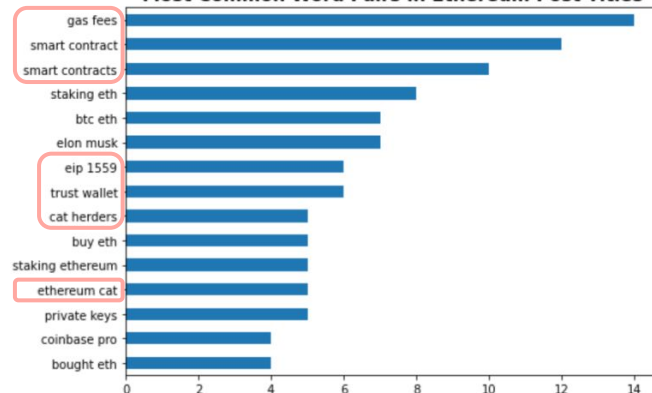
Top Word Pairs: Post Title

Most Common Word Pairs in Bitcoin Post Titles

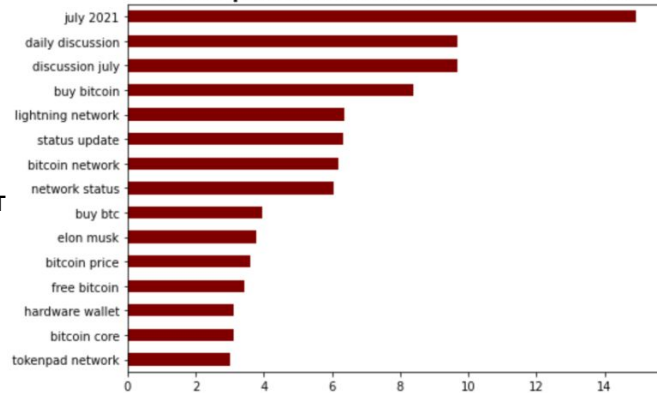


MOST
COMMON

Most Common Word Pairs in Ethereum Post Titles

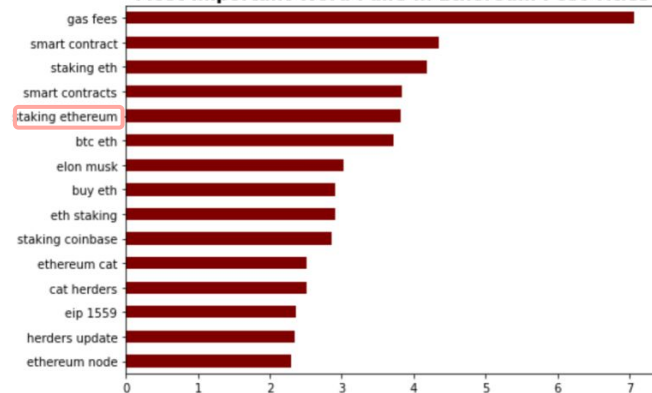


Most Important Word Pairs in Bitcoin Post Titles



MOST
IMPORTANT

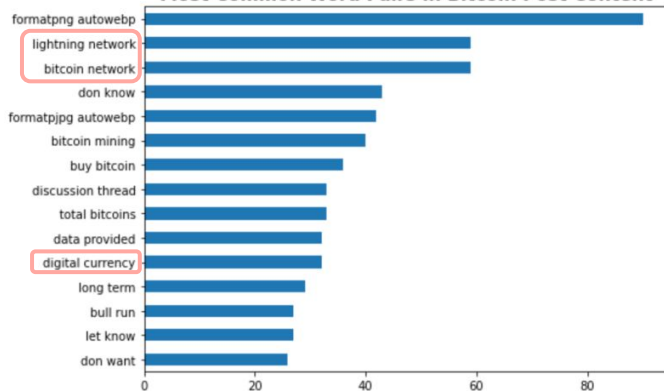
Most Important Word Pairs in Ethereum Post Titles



Note: Word importance was determined based on calculated TF-IDF

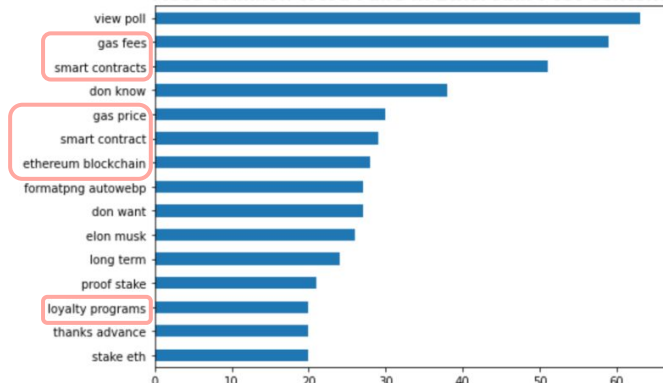
Top Word Pairs: Post Content

Most Common Word Pairs in Bitcoin Post Content

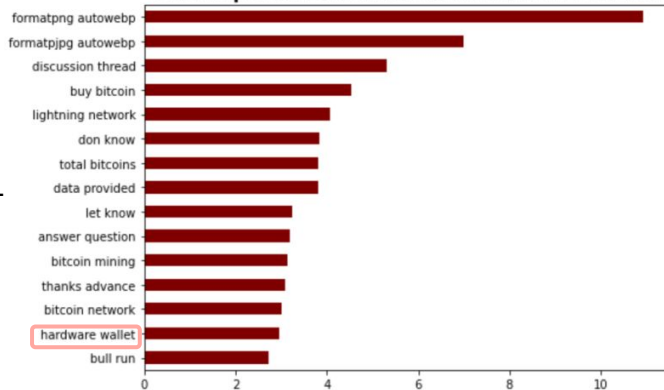


MOST
COMMON

Most Common Word Pairs in Ethereum Post Content

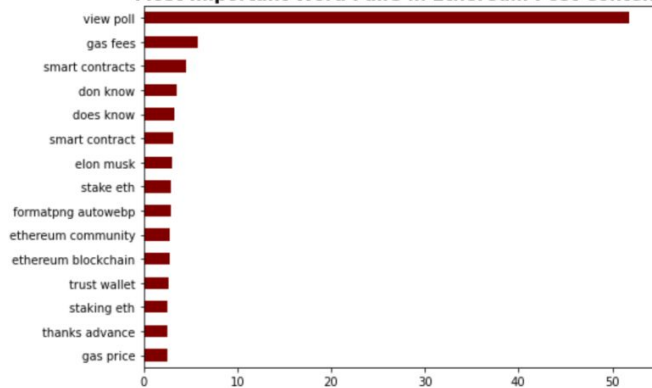


Most Important Word Pairs in Bitcoin Post Content



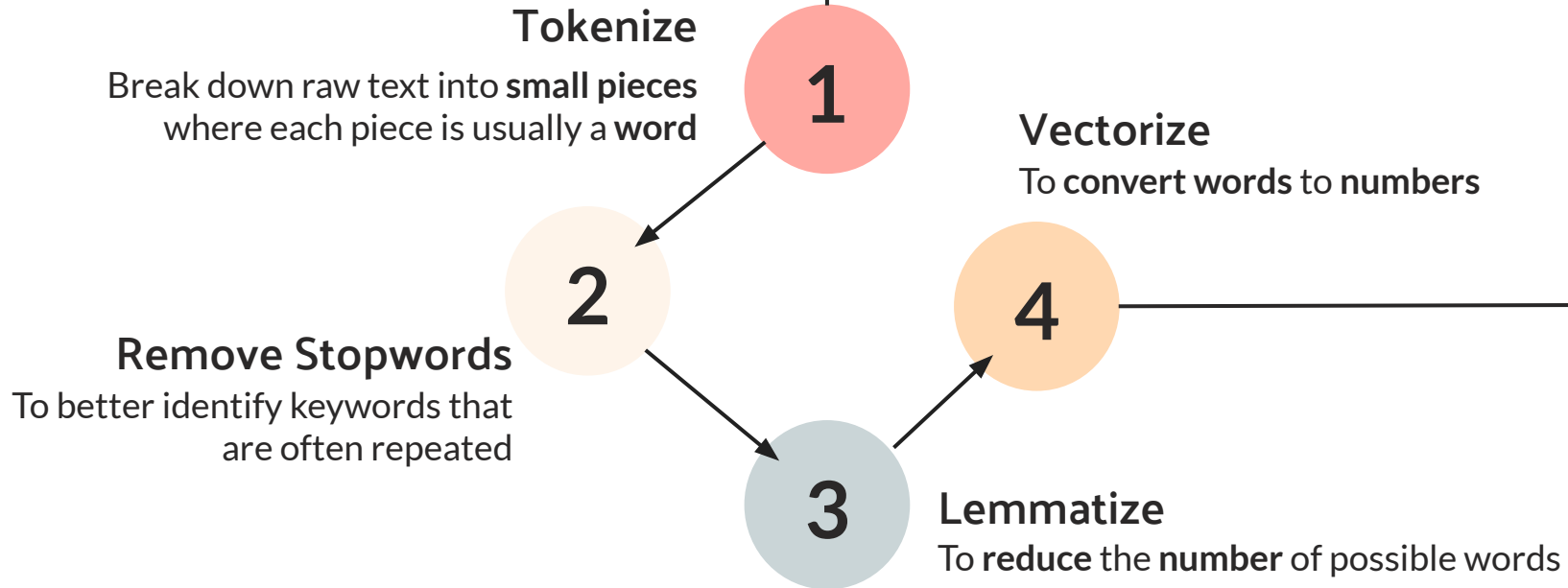
MOST
IMPORTANT

Most Important Word Pairs in Ethereum Post Content



Note: Word importance was determined based on calculated TF-IDF

Data Preprocessing



Note:

- Stopwords are words which do not add meaning to a sentence and can be removed without sacrificing the meaning of the sentence.
- Lemmatization is the reduction of reducing words to their actual root word (for e.g wolves -> wolf, sharing -> share)

Choice Of Vectorizers / Hyperparameters

Count Vectorizer

Chosen Hyperparameters for GridSearchCV

Max Features = 1000, 3000

N-Gram Range = (1, 1), (2, 2)

Min Document Frequency = 2

Max Document Frequency = 0.9

Tfidf Vectorizer

Choice Of Models / Hyperparameters

**Multinomial
Naive Bayes**

Chosen Hyperparameters for GridSearchCV

None

**K-Nearest
Neighbours**

Chosen Hyperparameters for GridSearchCV

No of Neighbours = 3, 5, 7, 9

Weights = Uniform, Distance

**Logistic
Regression**

Chosen Hyperparameters for GridSearchCV

C = 0.1, 1, 10

Solver = Liblinear

Penalty = L2, L1

Max Iterations = 10000

Performance Metrics

Accuracy

- Ranges from 0 to 1
- Measures how many **correct predictions** the model made out of **all the data points**

ROC AUC*

- Ranges from 0.5 to 1
- Quantifies how **well separated** the **underlying prediction distributions** made by the model are

*Receiver Operating Characteristic Area Under Curve

Vectorizer-Model Evaluation

Tfidf Vectorizer-
Logistic Regression

Cross Validation Accuracy

Count Vectorizer-
K-Nearest Neighbours

Train-Test Accuracy Difference

Tfidf Vectorizer-
Logistic Regression

Testing Accuracy

Tfidf Vectorizer-
Logistic Regression

Cross Validation ROC AUC

Tfidf Vectorizer-
Logistic Regression

Train-Test ROC AUC Difference

Tfidf Vectorizer-
Logistic Regression

Testing ROC AUC

Best Vectorizer-Model



**Tfidf
Vectorizer-
Logistic
Regression**

Best Hyperparameters from GridSearchCV

Max Features = 1000
N-Gram Range = 1, 1
Min Document Frequency = 2
Max Document Frequency = 0.9

C = 1
Solver = Liblinear
Penalty = L2

Baseline Model vs Chosen Model

57.5%

Baseline Accuracy

88.4%

Tfidf-Logistic Testing Accuracy

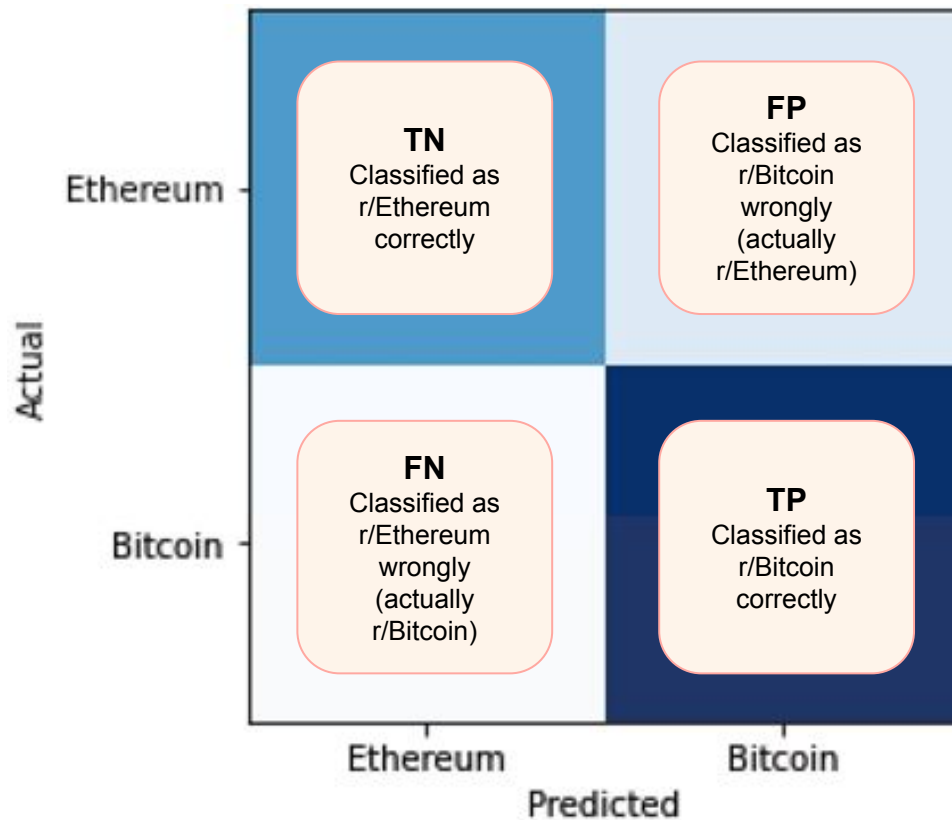
0.50

Baseline ROC AUC

0.95

Tfidf-Logistic Testing ROC AUC

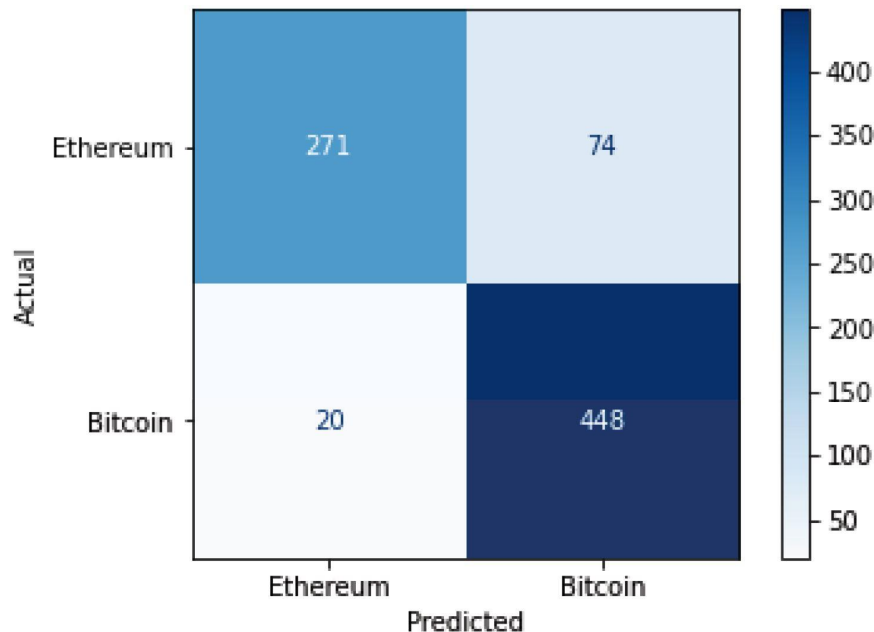
Demystifying The Confusion Matrix



Performance Metrics

On testing, our selected model:

- Classified submissions in the correct subreddits 88.4% of the time (also Accuracy)
- Precision = 85.8%
- Recall (also Sensitivity) = 95.7%
- F1 Score = 90.4 %



Total predictions = 271 + 74 + 20 + 448 = 813

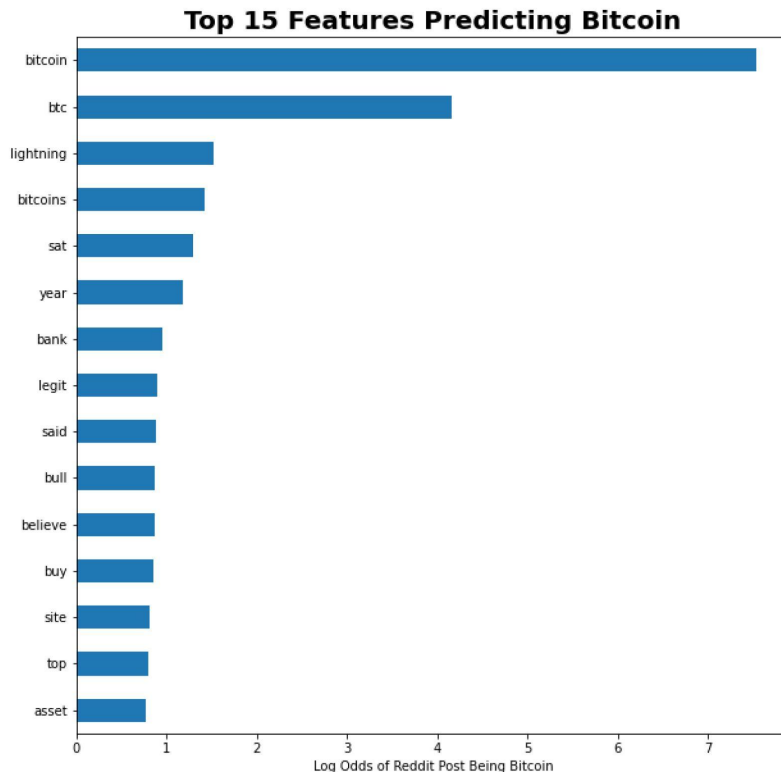
F1 Score = $2 * \text{Recall} * \text{Precision} / (\text{Recall} + \text{Precision})$; provides a better overall measure of performance

Precision measures how precise a classifier is out those predicted positive, how many of them are actual positive.

Recall measures how good a classifier is at detecting positives.

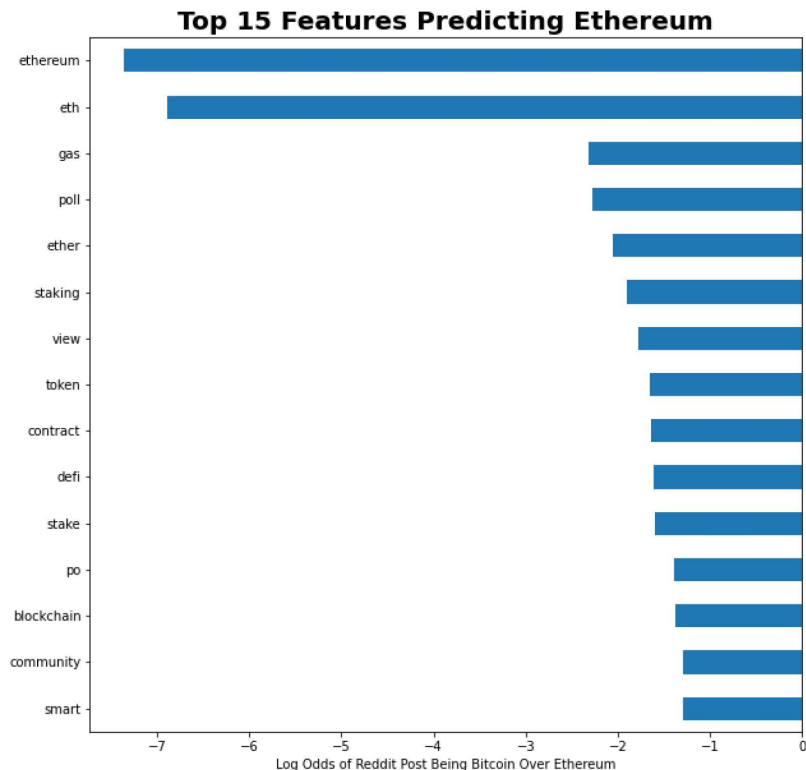
Top Features For r/Bitcoin

- Expected to find *bitcoin* and *btc* at the top of the list
- Most words except for 'lightning' do not associate well with Bitcoin
- 'lightning' is likely linked to Lightning Network which is a layer payment protocol designed to be layered on top of a blockchain-based cryptocurrency such as bitcoin or litecoin



Top Features For r/Ethereum

- Expected to find *ethereum* and *eth* at the top of the list
- More associative words like *gas*, *ether*, *staking*, *smart*, *defi* and *contract*, which are specific to Ethereum
- They refer to different aspects of what goes under the hood in Ethereum's blockchain protocol and its applications



Chatbot Demo



CryptoTrade23Bot

July 30

Good morning! Ask me a question about cryptocurrency

What is Bitcoin?

Bitcoin is a decentralized digital currency, without a central bank or single administrator, that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries.

Write a message



Chatbot Demo



CryptoTrade23Bot

July 30

Stake ether?

Do you mean Ethereum staking?

Yes

Staking is the act of depositing 32 ETH to activate validator software. This will keep Ethereum secure for everyone and earn you new ETH in the process.

Write a message

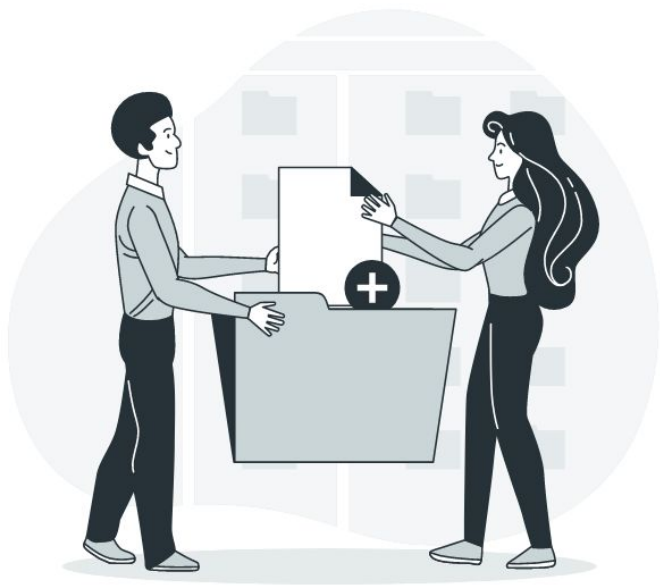


Conclusion

- MVP chatbot should be piloted to a targeted group of users of our platform to determine if it is effective and productive in classifying and responding to text enquiries accurately
- Building an ensemble model that combines different models (e.g. Naive Bayes, LogReg and/or Support Vector Machine) can deliver a more optimal performance
- One potential future development is the expansion in the scope of our algorithm to include more cryptocurrencies by extending the existing binary classifiers to multi-class classifiers
- Another potential development is use of sentiment analysis to augment our core text classification

**Thank
You**





Appendix

Term Frequency - Inverse Document Frequency (TF-IDF)

- reflects how important a word is to a document, in a collection

Term Frequency

- No. of **instances** of a term in a **single** document

No. of **instances** of a term in a
single document

No. of words in the
document

X

Inverse Document Frequency

- The more **unique** a term is to a document, the higher the **weight** assigned
- For e.g “the” vs “ethereum”

log $\frac{\text{No. of documents}}{\text{No. of documents containing the term}}$

Vectorizer-Model Evaluation

No	Transformer	Estimator	CV Accuracy	CV ROC AUC	Training Accuracy	Testing Accuracy	Accuracy Difference	Training ROC AUC	Testing ROC AUC	ROC AUC Difference
1	Count Vectorizer	Multinomial Naive Bayes	0.852	0.919	0.923	0.859	0.064	0.977	0.920	0.057
2	Tfidf Vectorizer	Multinomial Naive Bayes	0.851	0.925	0.892	0.845	0.047	0.964	0.927	0.037
3	Count Vectorizer	K-Nearest Neighbours	0.763	0.865	0.854	0.845	0.009	0.931	0.860	0.071
4	Tfidf Vectorizer	K-Nearest Neighbours	0.803	0.872	0.857	0.808	0.049	0.933	0.876	0.057
5	Count Vectorizer	Logistic Regression	0.869	0.942	0.971	0.881	0.090	0.997	0.944	0.053
6	Tfidf Vectorizer	Logistic Regression	0.869	0.949	0.939	0.884	0.055	0.985	0.950	0.035