Web Scraping & NLP

Bitcoin & Dogecoin Subreddits

BY: OLIVIA LARA

OVERVIEW

Table of Contents

01 Problem Statement

02 Data Info

03 Data Cleaning

04 Exploratory Data Analysis

05 Models & Scores

06 Conclusion

PROBLEM STATEMENT

Are we able to classify whether a given post belongs to -r/Bitcoin or -r/Dogecoin?

Which features impact our classification model and what do these tell us about Reddit Bitcoin users and Reddit Dogecoin users?

BASIC INFO ON DATA

- Subreddit Data
    - Bitcoin and Dogecoin
    - Year: 2019 - 2021
- Data
    - 2,600 Observations
    - 11 Features (2 Used in Model)

METHODOLOGY

01 **Collected & Cleaned Data**

Web Scraping, Empty Values, Emoji to Text, & Regex

02 **Exploratory Data Analysis**

Word Counts, Distribution & Popularity of Words & Emojis

03 **Models**

7 Different Models, Transformers, Pipelines, & GridSearchCV

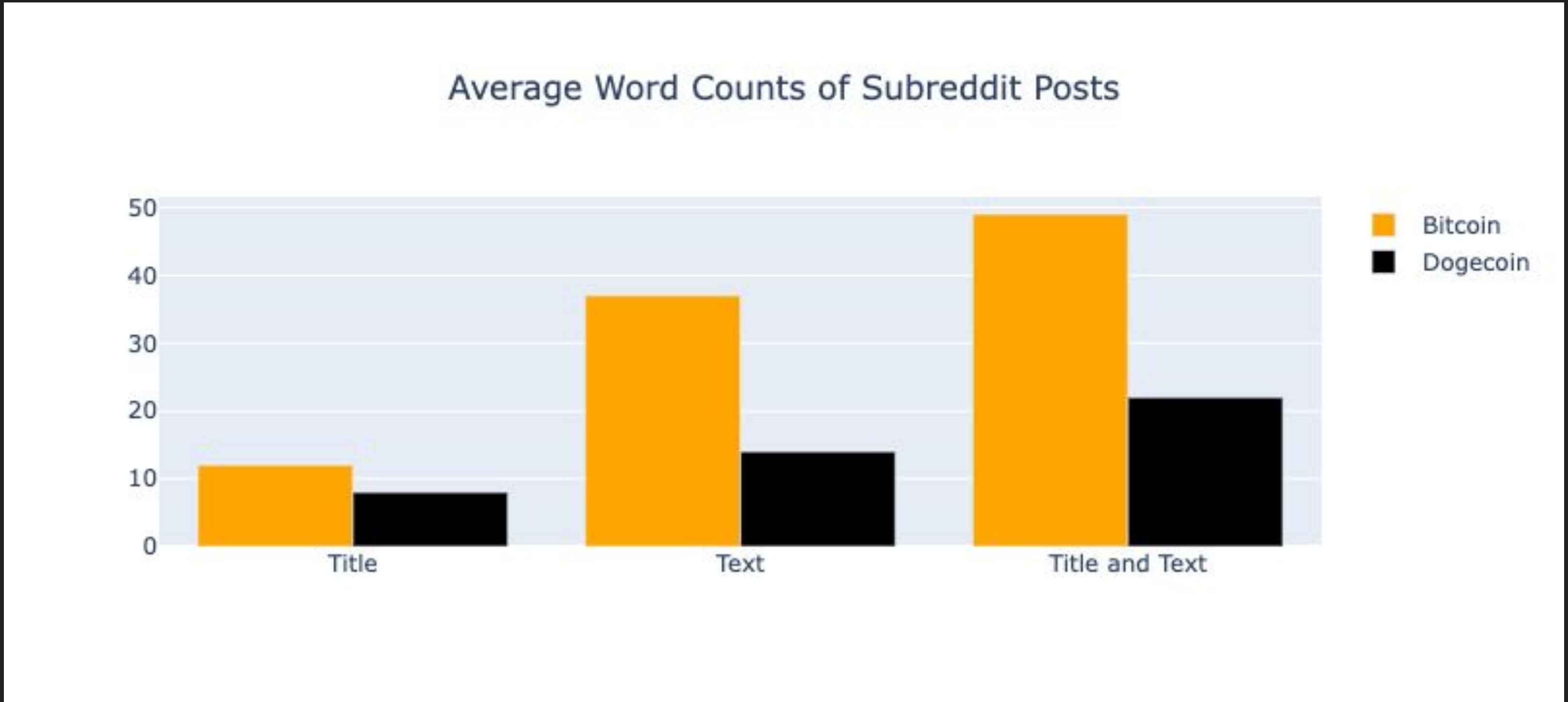
04 **Scores and Coefficients**

Accuracy Score and Feature Importances

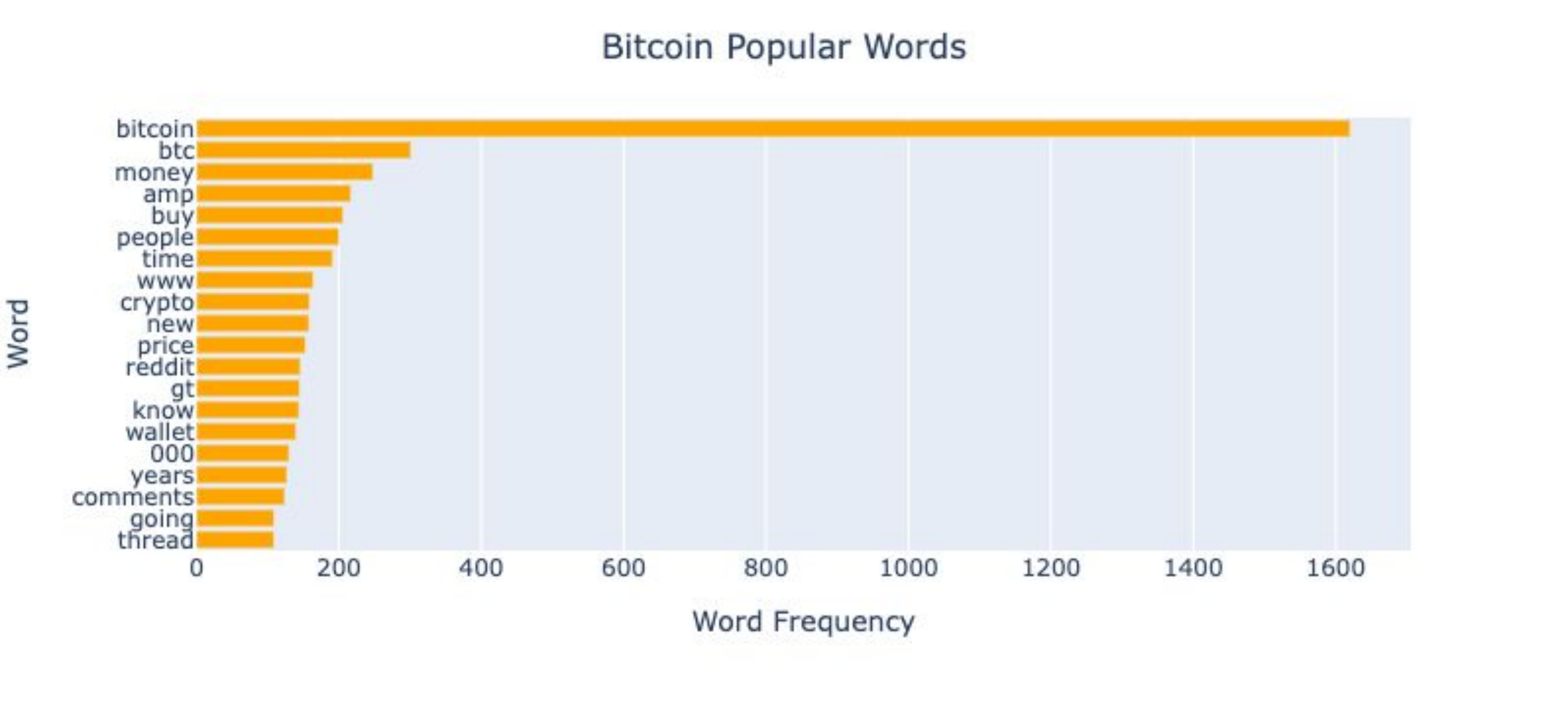
CLEANING DATA

- Started with 3 features: title, text, year
- Dealt with NaN values
- Added features: title and text, text word count, title word count, total word count, emoji column, emoji to text column, and emoji count
- Removed unnecessary characters with RegEx
- Ended with 11 features (column names)

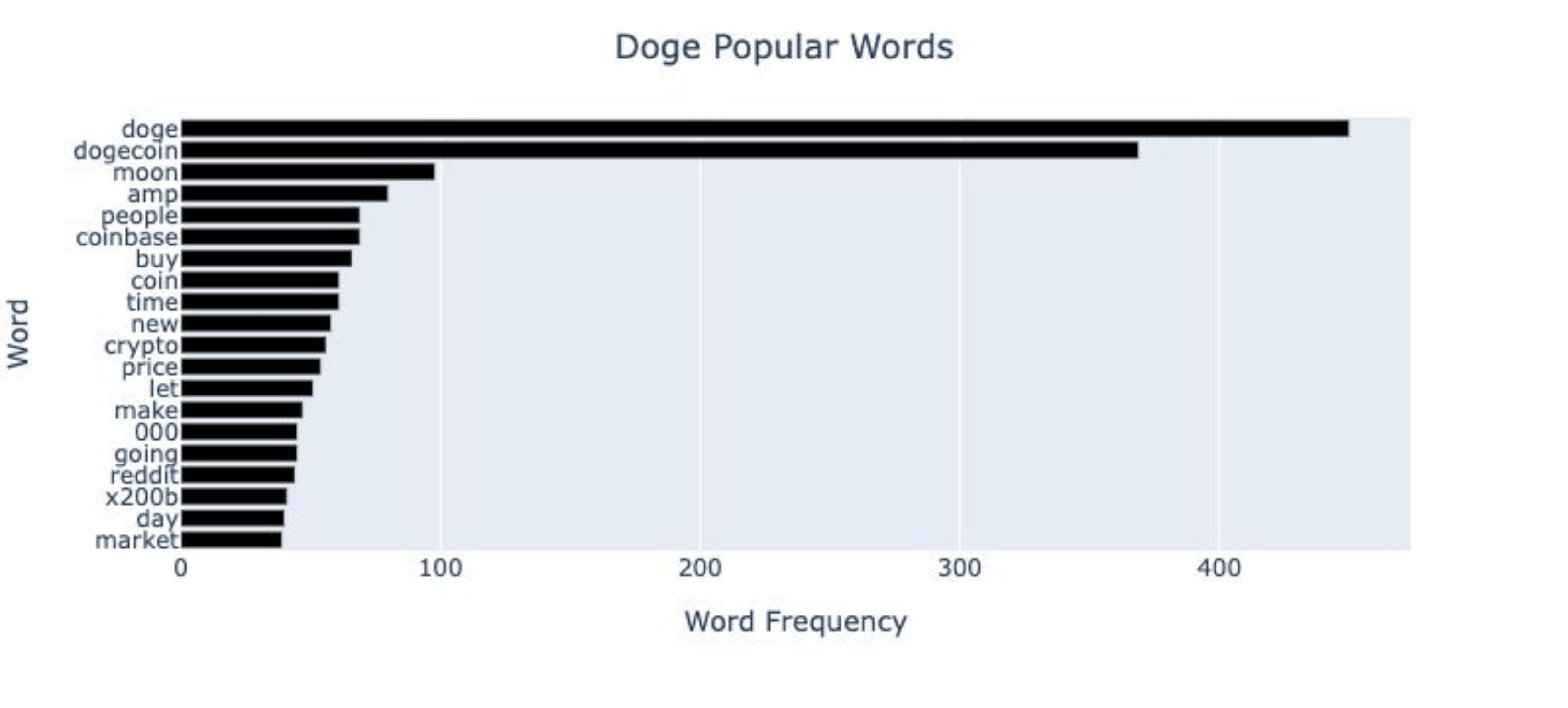
AVERAGE WORD COUNTS OF POSTS



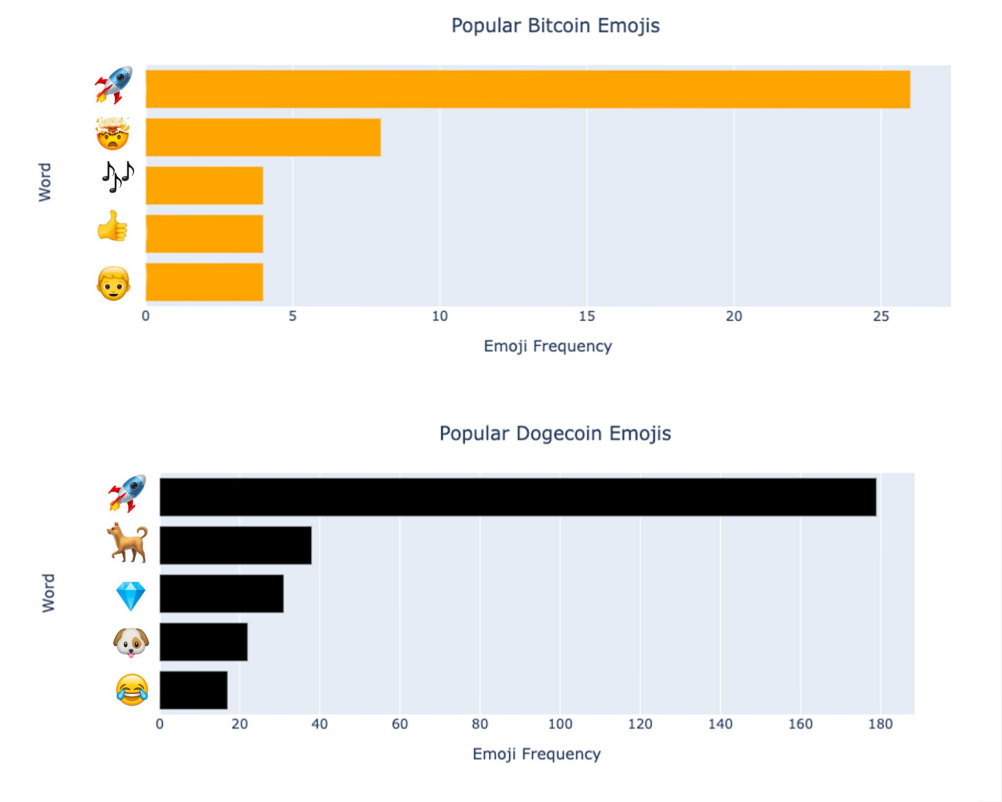
POPULAR WORDS USED IN POSTS



POPULAR WORDS USED IN POSTS



POPULAR EMOJIS USED IN POSTS



Popular Bitcoin Emojis

Popular Dogecoin Emojis

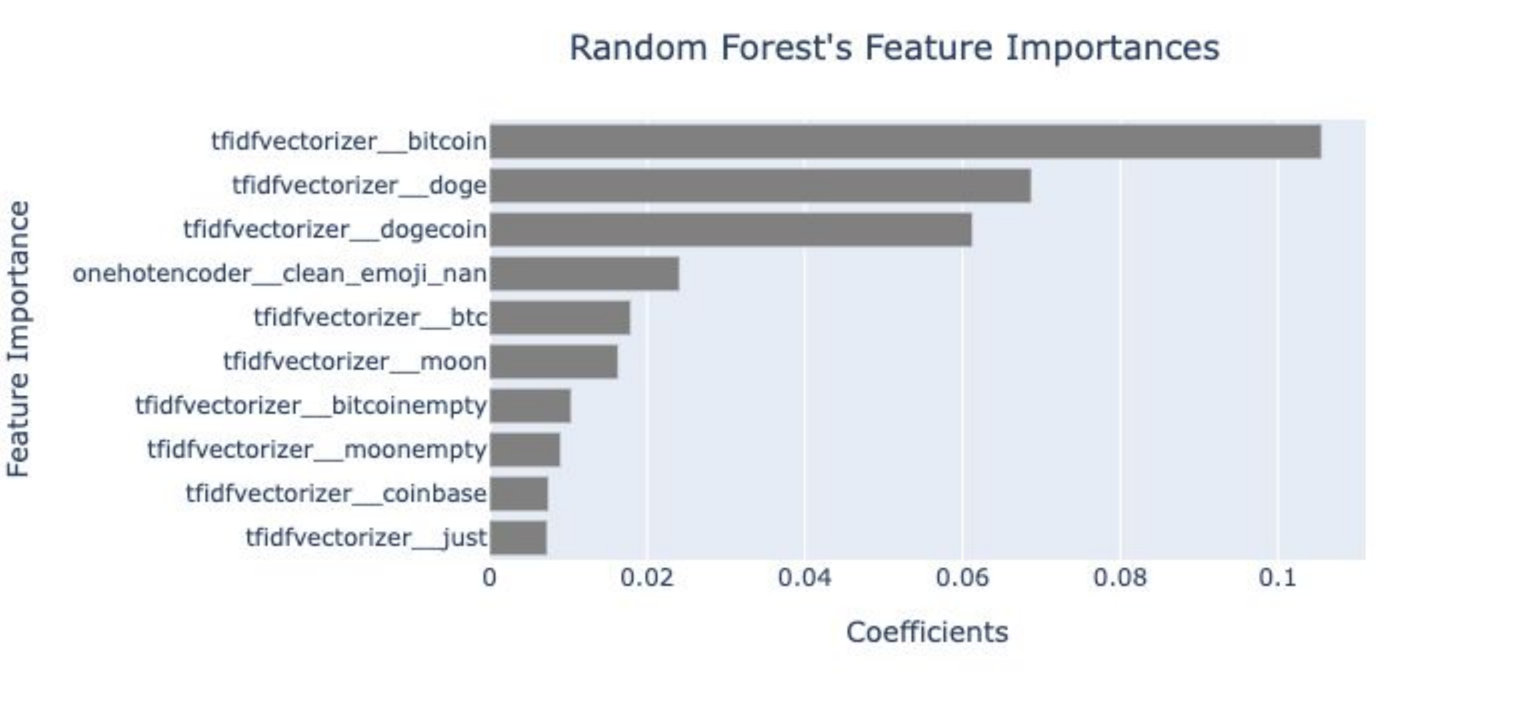
MODEL SCORES

Model	Transformer	Train Score	Test Score
Null Model	N/A	.616	.616
Logistic Regression	Count Vectorizer	.989	**.885**
KNeighbors Classifier	Tfidf Vectorizer	.885	.805
Decision Tree Classifier	Count Vectorizer	.999	**.888**
Bagging Classifier	Tfidf Vectorizer	.993	**.895**
Random Forest Classifier	Count Vectorizer	.999	**.883**
Ada Boost Classifier	Count Vectorizer	.892	.863
Gradient Boosting Classifier	Tfidf Vectorizer	.901	.866

BEST MODEL: RANDOM FOREST

- Grid Search
    - stop _words: “english”
    - n_gram_range: (1,1)
    - n_estimators: 1,000
    - warm_start: false
- Scores:
    - Train: 0.998
    - Test: 0.9

FEATURE IMPORTANCES



CONCLUSIONS

01

Random Forest is Best Estimator

Test Score: .900

03

Dogecoin Reddit Users Love Emojis & Memes

Emojis used in 20% of posts and moon is a very impactful word

SOURCES

- https://www.reddit.com/r/dogecoin/
- https://www.reddit.com/r/Bitcoin/
- https://api.pushshift.io/reddit/search/submission?subreddit=Bitcoin
- https://slidesgo.com/theme/stds-diseases-awareness-brochure#search-Editor%27s+Choice&position-13&results-1594

THANK YOU FOR LISTENING!