Bitcoin Tweet Sentiment Analysis

Natalya Doris
Flatiron School, Data Science

Phase 4

Objective

This project seeks to build a model that **accurately classifies** tweets about Bitcoin as having either positive or negative sentiment. Unlabeled tweets classified by this model could ultimately could be used to analyze time trends on Bitcoin sentiment and assess the predictive power of Twitter sentiment on future price movements of the cryptocurrency.

Breaking it Down

Explore the Data

- Understand **trends** and key **takeaways**
- **Clean** the text data so that it is ready to be used in modeling

Build a Model

Test and **tune**:

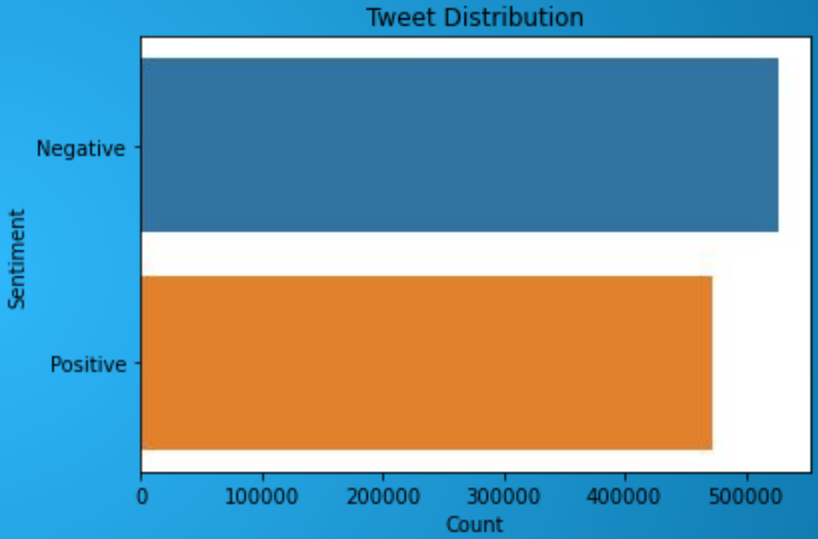
- Three vectorizers
- Four classification models

Evaluate the Model

Score each model, selecting a final **best-performing** model

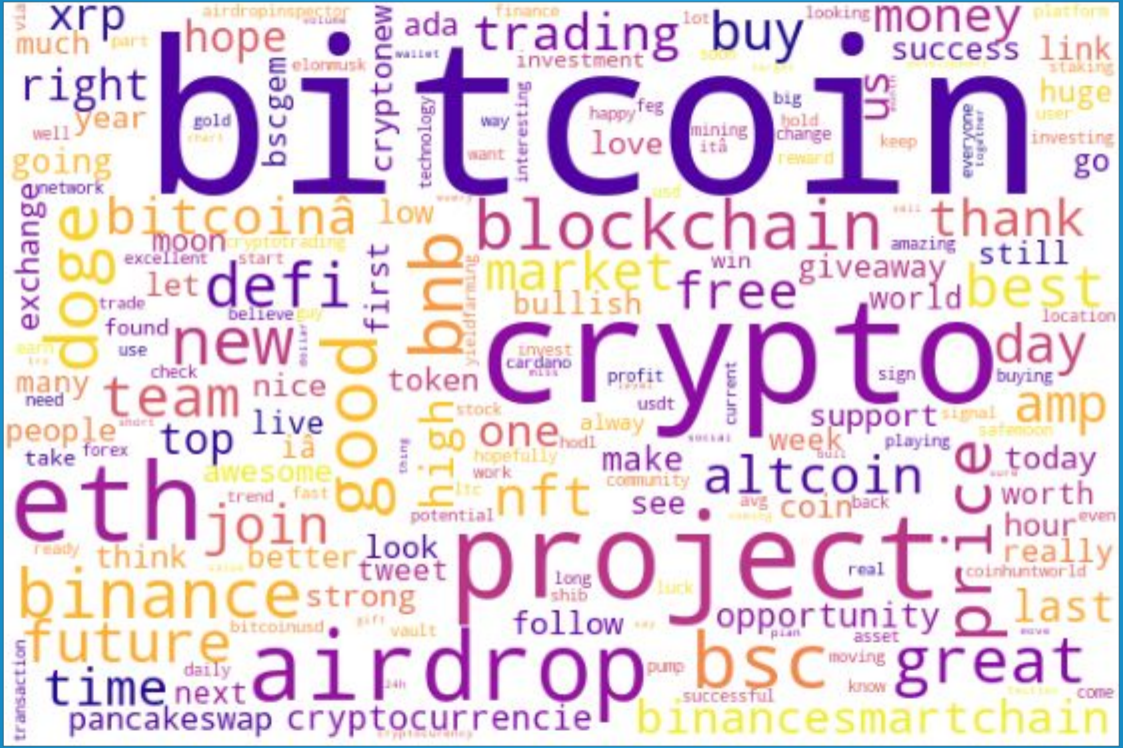
The Data:

One million Tweets referencing Bitcoin, spanning a six-month period from February 2021 to August 2021

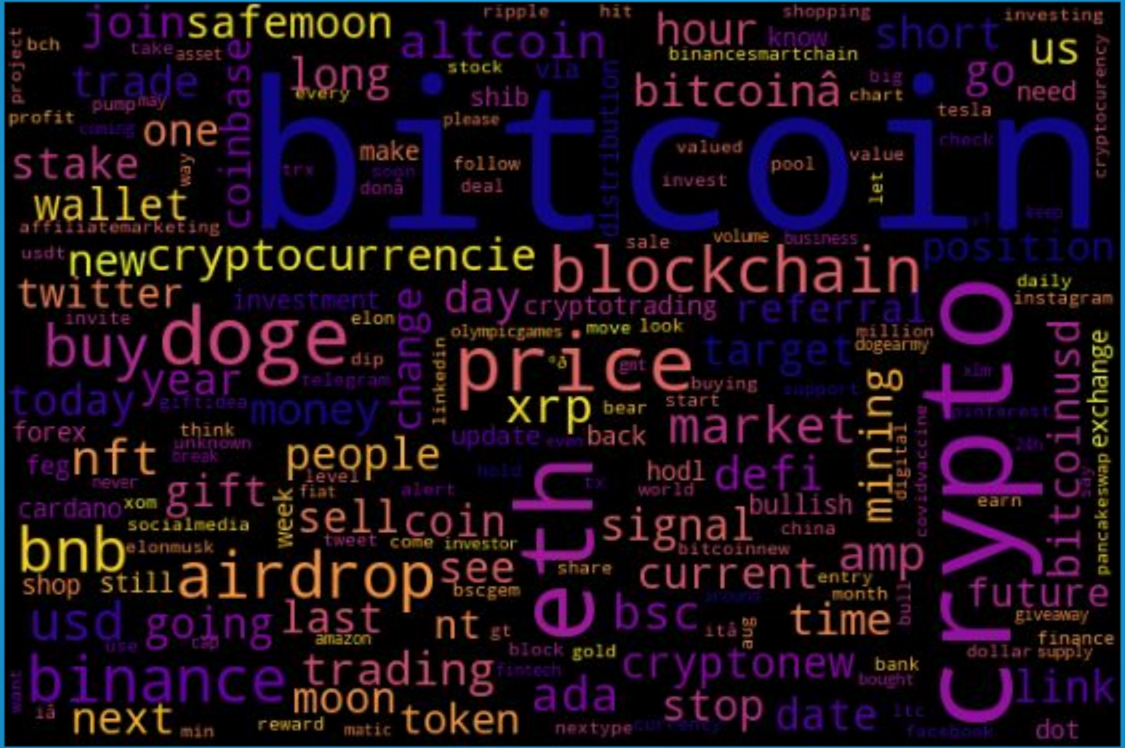


Source: Kaggle

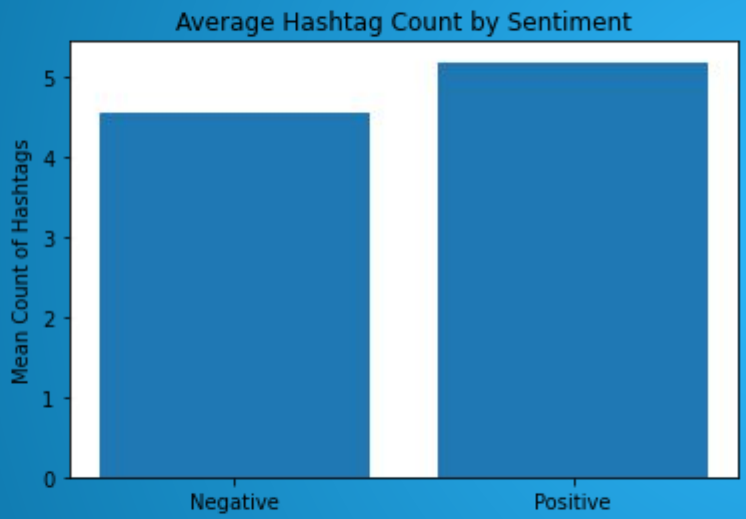
Word Cloud of Positive Tweets

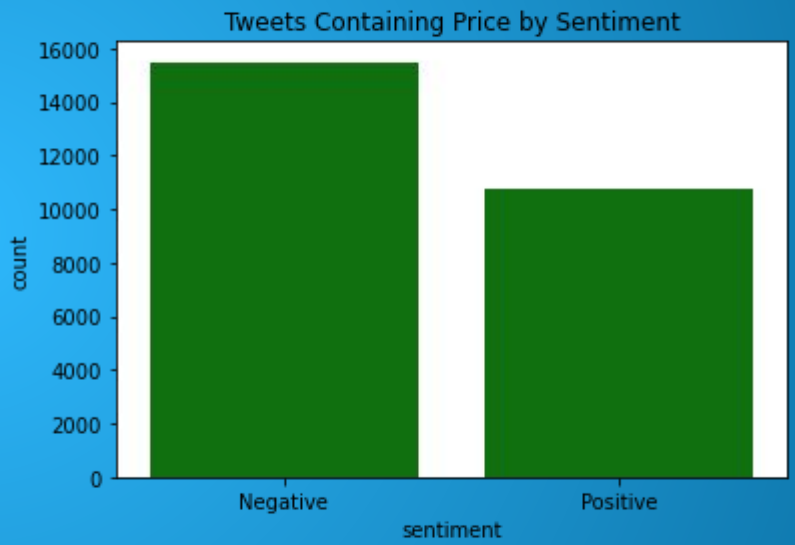


Word Cloud of Negative Tweets



Relevant Trends

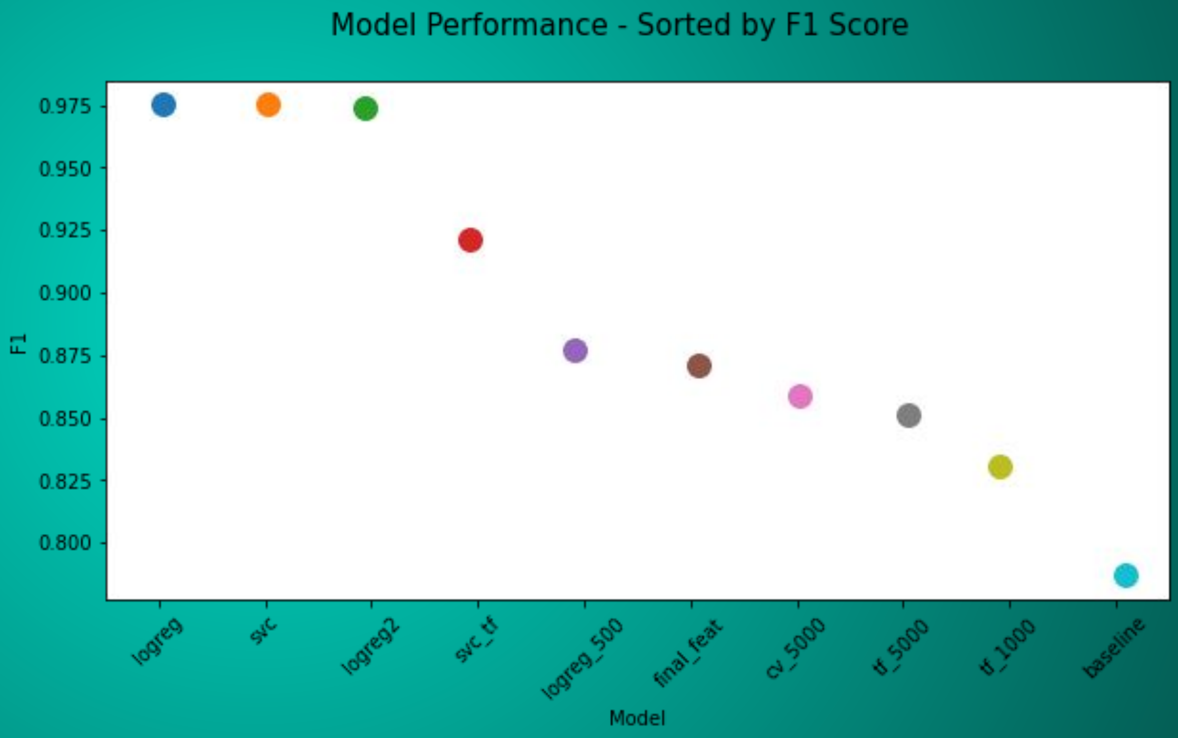




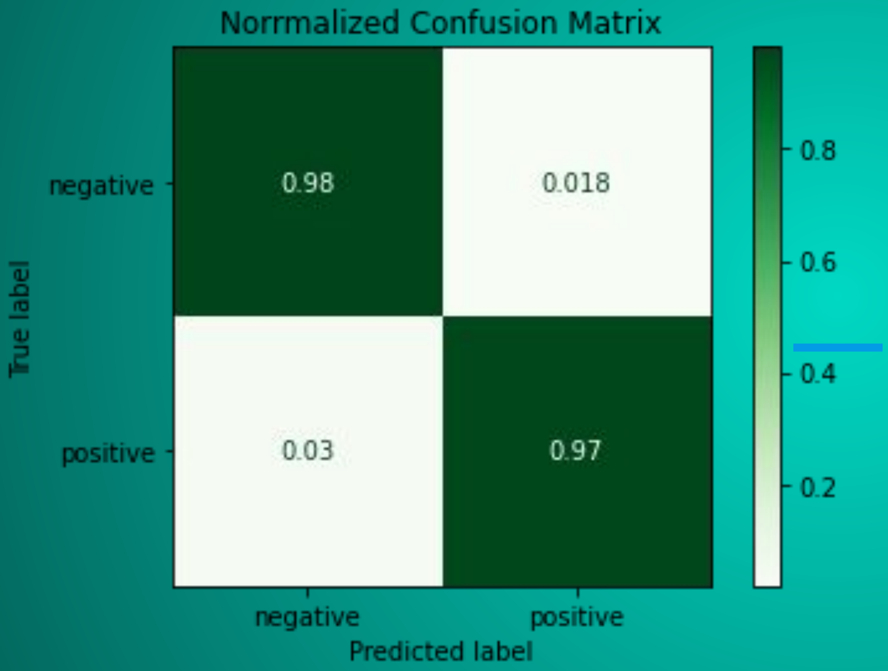
- Positive tweets had more hashtags on average, negative tweets more frequently contained a price

Modeling

- Final model is a Logistic Regression Classifier with a Count Vectorizer
- High F1 score indicates model both captures positive cases (recall) and is accurate with the cases it does capture (precision)
- Final model is a ~20% improvement from the baseline



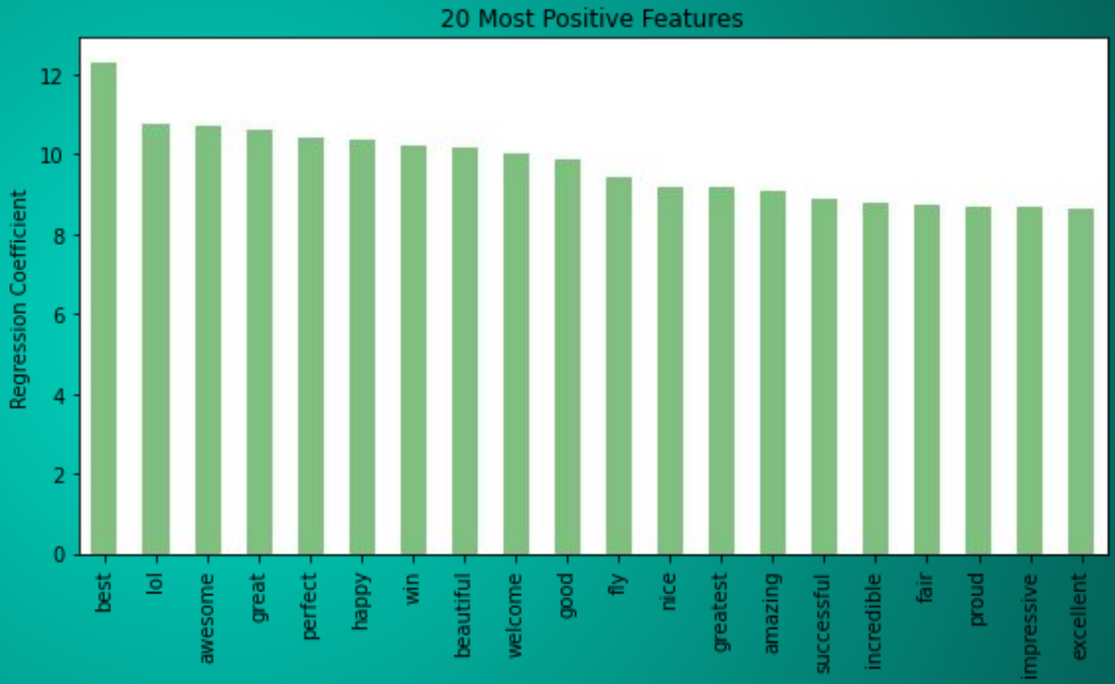
Modeling, Part II



- Final model is **97% accurate** overall
- Just **3%** of validation data categorized as negative when it was actually positive
- Just **1.8%** of validation data categorized as positive when it was actually negative

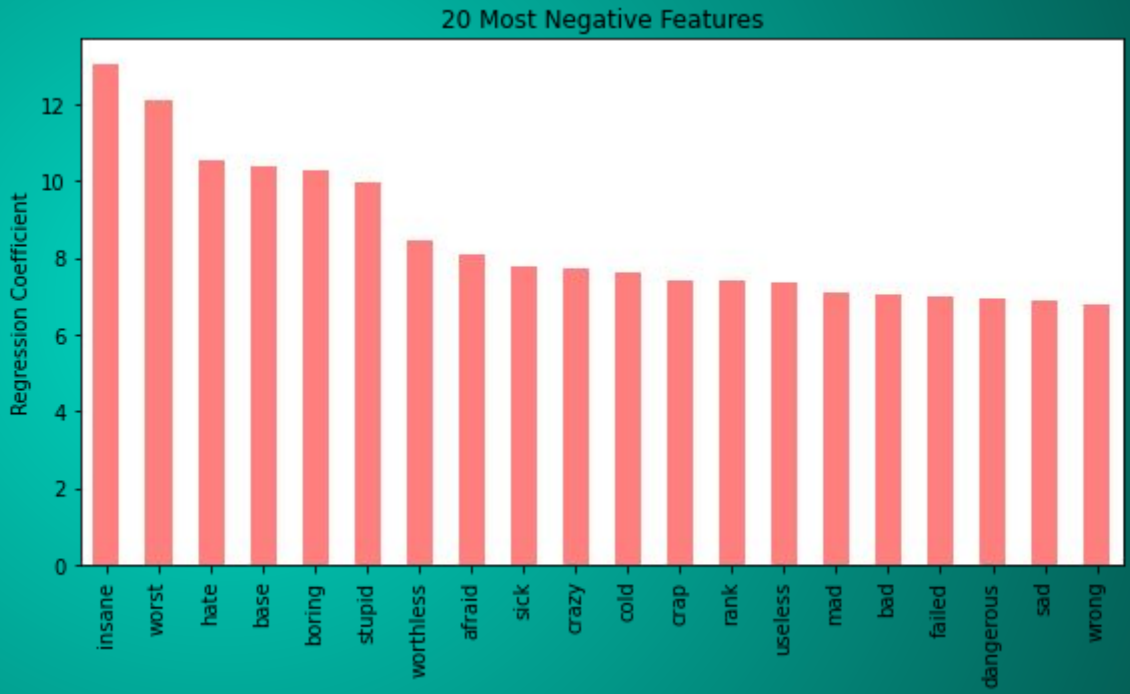
Feature Importance

Words like **‘best’**, **‘awesome’**, **‘successful’** of highest importance in predicting positive sentiment



Feature Importance

Words like ‘**insane**’, ‘**worst**’, ‘**worthless**’ of highest importance in predicting negative sentiment



Conclusions

Model is more than **40% better** at classifying sentiment than random guessing

- A Logistic Regression model was the best-performing classifier, with Count Vectorization used to process the annotated tweets
- 97% accuracy, 97% F1 score indicates model captures positive cases (recall) without casting too wide a net, i.e. little misclassification in either direction (precision)

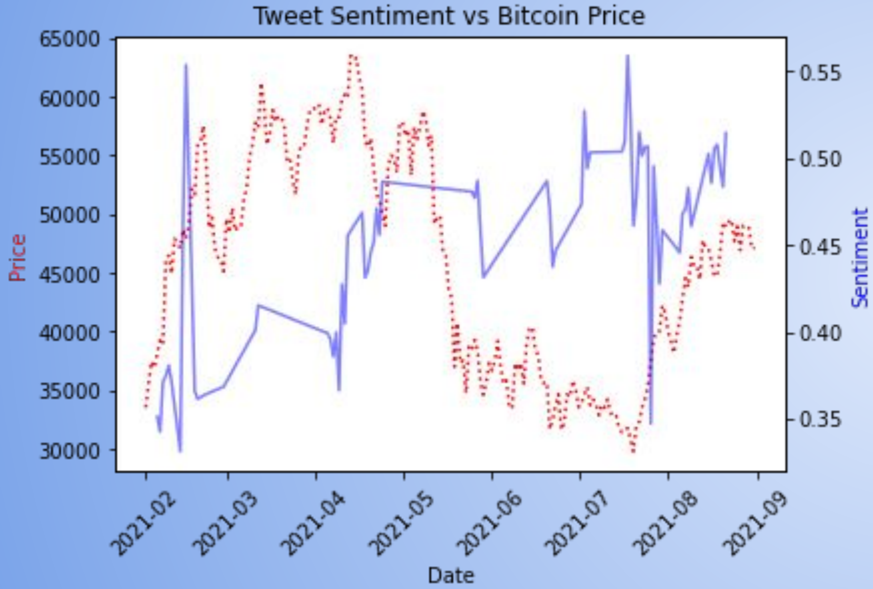
Digging Deeper

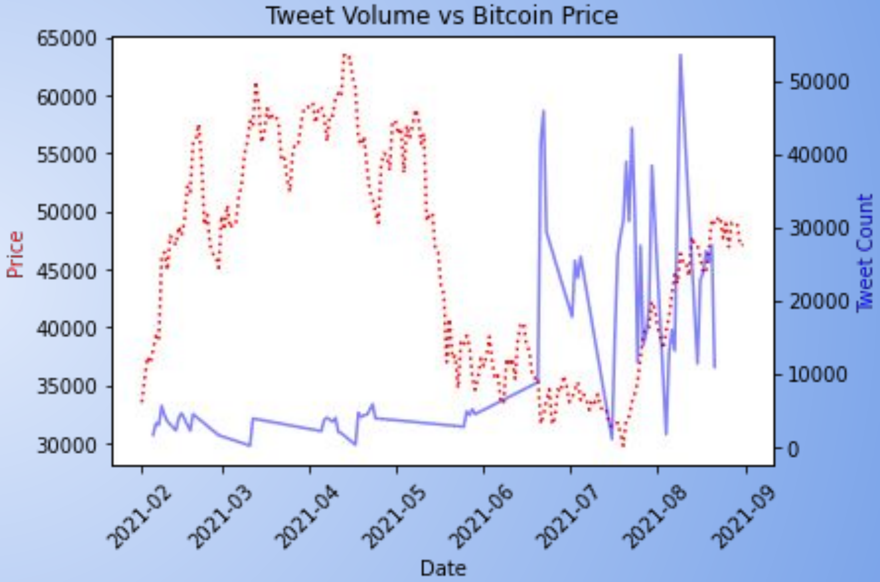
- Words important to the model included ‘best’, ‘awesome’, ‘successful’, ‘insane’, ‘worst’, ‘worthless’
- Positive tweets had more hashtags on average, negative tweets more frequently contained a price

Next Steps & Recommendations

1. *Run the model on real-time Tweets about Bitcoin, pulled via Twitter API*
2. *Use model-labeled Tweets to conduct Time Series Analysis, with the aim of understanding the predictive power of Tweet sentiment on the price of BTC*

A Preview: Time Series





Thank you!

Contact Info:

- ntdoris2@gmail.com
- https://github.com/ntdoris