Bitcoin Tweet Sentiment Analysis

By Natalya Doris
Flatiron School, Data Science
Phase 4

Business Problem

In this project, we help Crypto Consultancy firm X build a model that can accurately **classify** Tweets about Bitcoin as having either positive or negative **sentiment**. Firm X would like to use the model to classify **unlabeled** Tweets and **understand** any characteristics distinguishing the positive and negative Tweets.

Breaking it Down

Explore the Data

- Understand **trends** and key **takeaways** to bring to the client
- **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

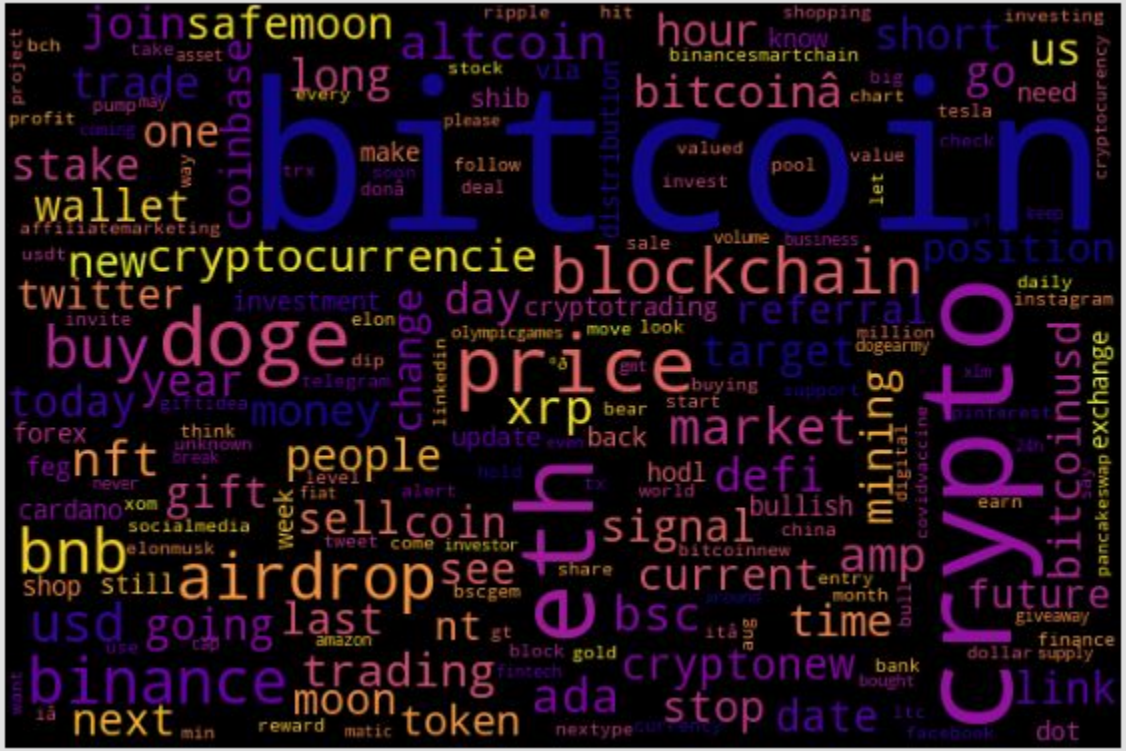
The Data:

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

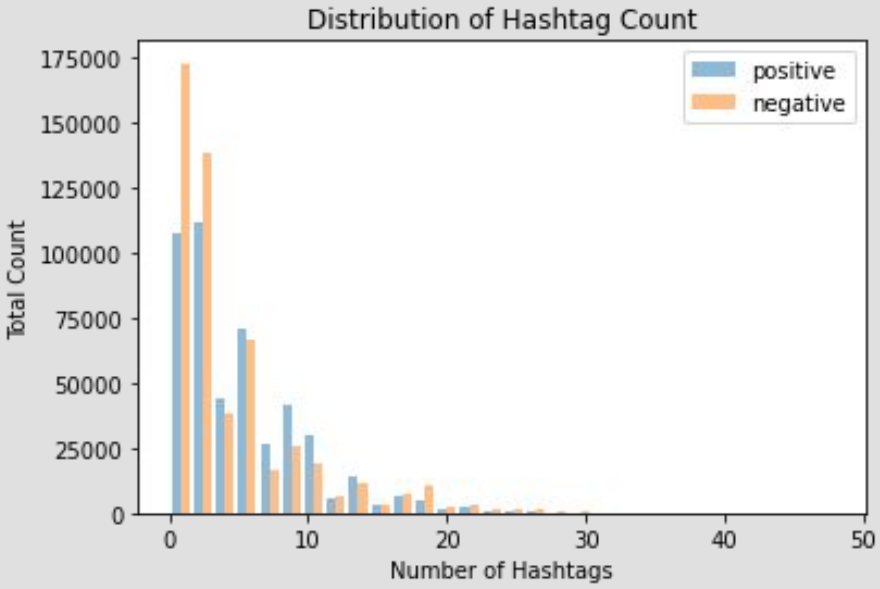
Source: Kaggle

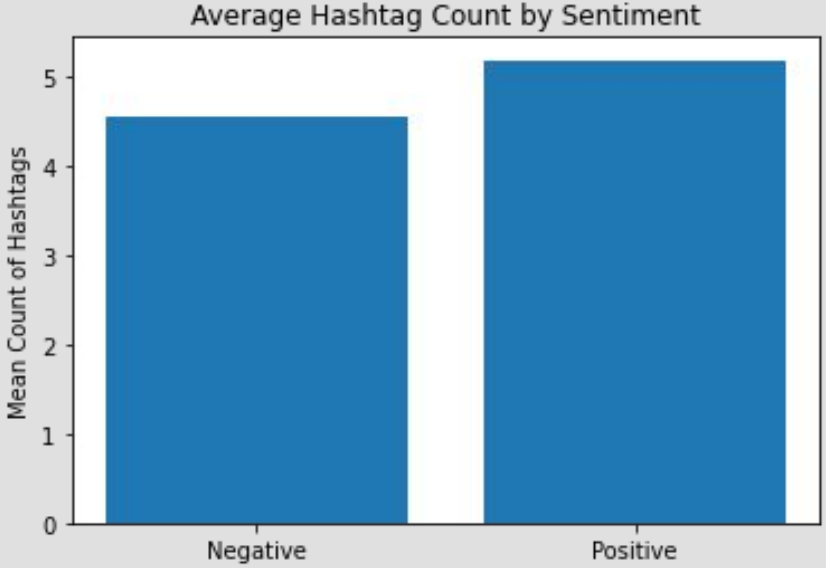
Word Cloud of Positive Tweets

Word Cloud of Negative Tweets

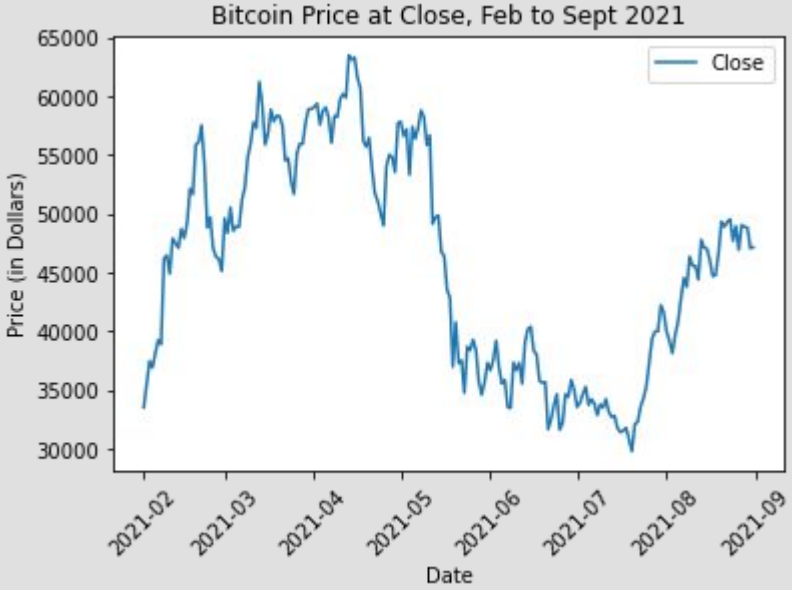


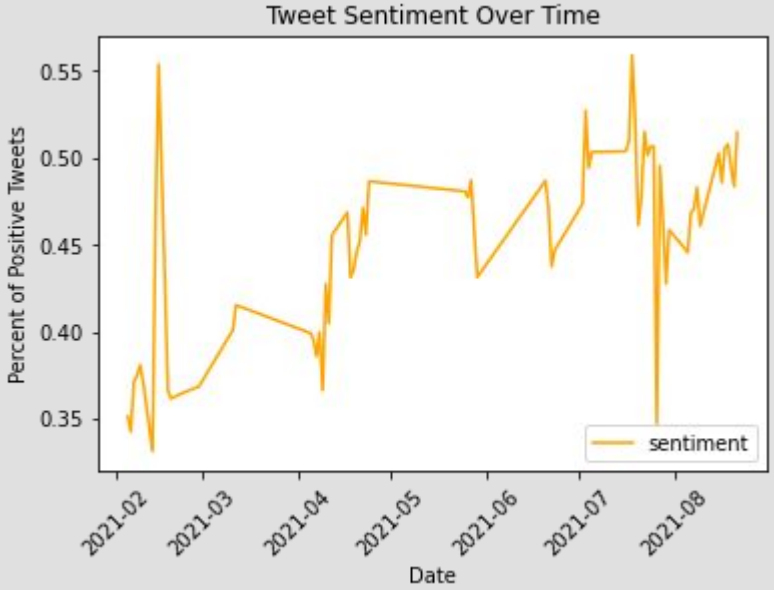
Hashtags


Distribution of Hashtag Count
positive
negative
Total Count
175000
150000
125000
100000
75000
50000
25000
0
0
10
20
30
40
50
Number of Hashtags

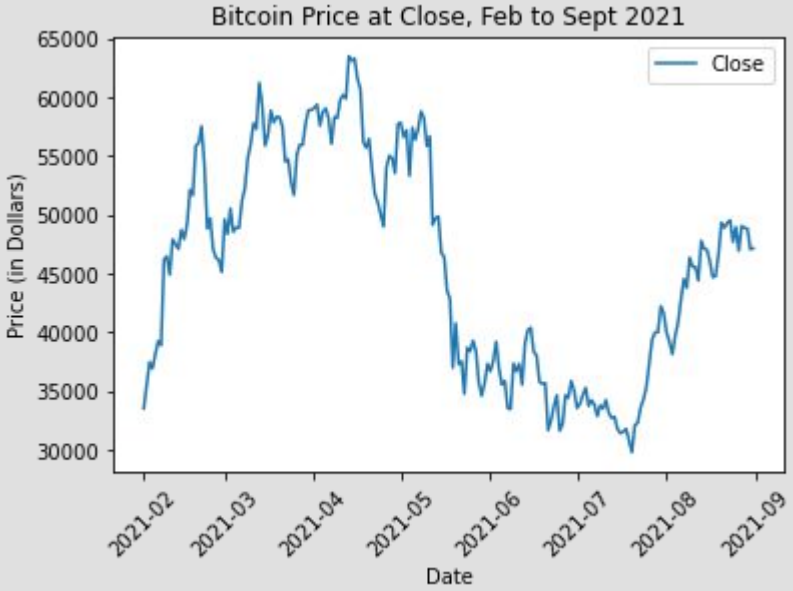

Average Hashtag Count by Sentiment
Mean Count of Hashtags
5
4
3
2
1
0
Negative
Positive

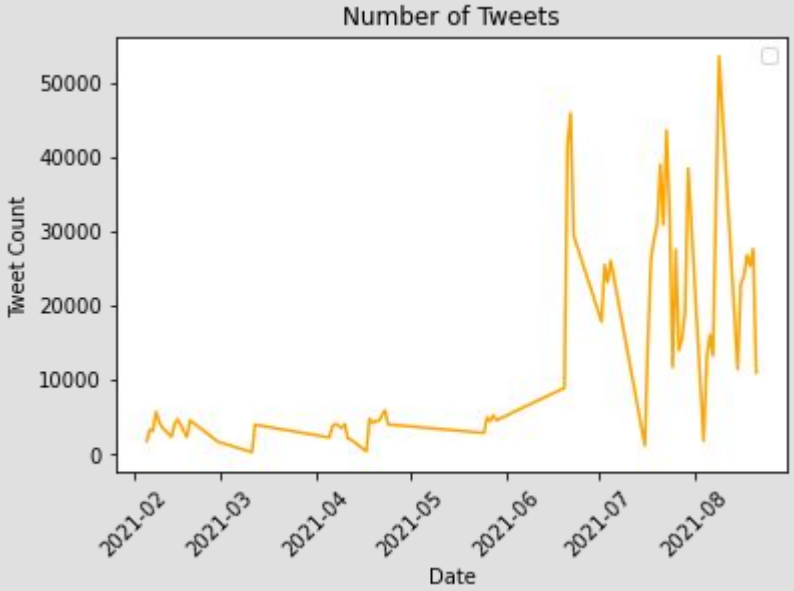
#Trending


Bitcoin Price at Close, Feb to Sept 2021
Close
Price (in Dollars)
65000
60000
55000
50000
45000
40000
35000
30000
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
2021-09
Date


Tweet Sentiment Over Time
Percent of Positive Tweets
0.55
0.50
0.45
0.40
0.35
sentiment
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
Date

#Trending


Bitcoin Price at Close, Feb to Sept 2021
Close
Price (in Dollars)
65000
60000
55000
50000
45000
40000
35000
30000
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
2021-09
Date


Number of Tweets
Tweet Count
50000
40000
30000
20000
10000
0
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
Date

Modeling

	f1_train	f1_test	accuracy_train	accuracy_test	roc_auc	pr_auc	model	vectorizer
baseline	0.575926	0.575227	0.619592	0.619370	0.615580	0.547573	MultinomialNB()	TfidfVectorizer(max_features=10)
base_tf	0.873472	0.853652	0.883644	0.865230	0.863466	0.809588	MultinomialNB()	TfidfVectorizer()
base_cv	0.967555	0.858882	0.969429	0.869114	0.867714	0.813043	MultinomialNB()	CountVectorizer(ngram_range=(2, 2),\n ...
base_cv2	0.958314	0.867798	0.960734	0.875402	0.874822	0.817498	MultinomialNB()	CountVectorizer(ngram_range=(2, 2))
logreg	0.991206	0.917310	0.991740	0.925155	0.922720	0.901446	LogisticRegression()	CountVectorizer(ngram_range=(2, 2))
svc	0.999980	0.926771	0.999981	0.933035	0.931121	0.909578	LinearSVC()	CountVectorizer(ngram_range=(2, 2))

Final model is Linear SV with Count Vectorizer, achieving ~93% accuracy and F1 score!

Conclusions

Conversation Matters	• Period of higher Twitter volume associated with period of lower Bitcoin price
Sentiment Impactful	• Spike in positive sentiment occurred around same time as rise in Bitcoin price
#Positive	• Positive sentiment Tweets tend to have more hashtags on average
Model Accuracy	• Final model can classify unlabeled Tweets as positive or negative with ~90% accuracy

Next Steps

- Pull more recent Tweets on Bitcoin via Twitter API
- Run final model on unseen data, i.e. new Tweets
- 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
- Use deep learning models to classify Tweets, comparing these results to previous model