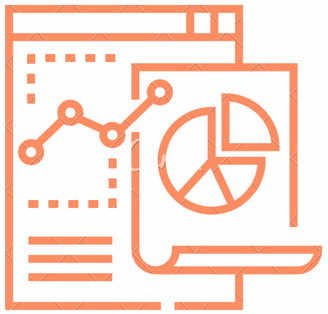
Bitcoin Price Forecasting and Sentiment Analysis using Tweets

Empirical Project

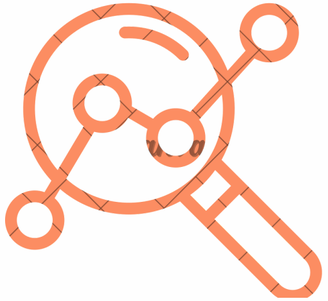
Giulia Luciani
Alessandro Sottile

TABLE OF CONTENTS



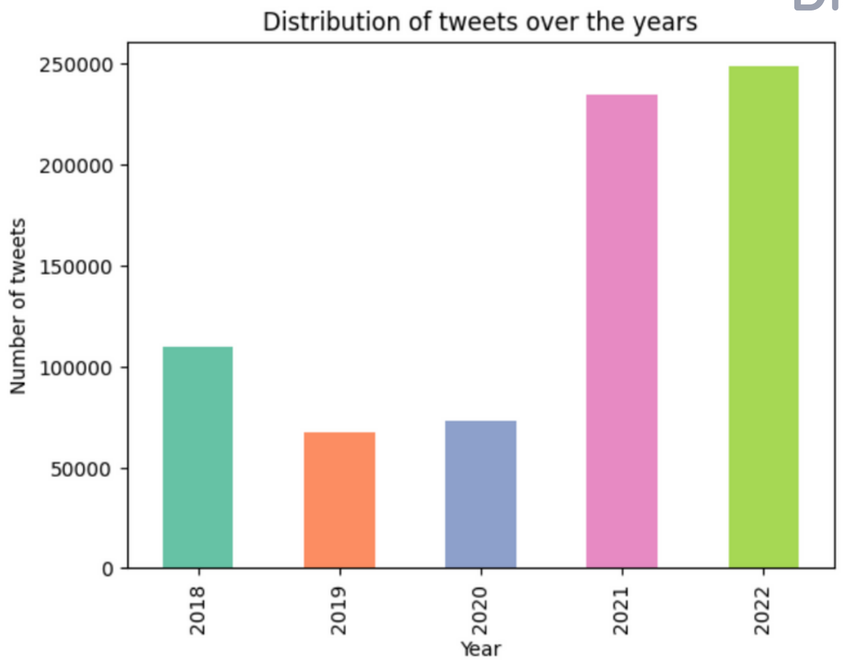
Exploratory Data Analysis

Leading Indicators



Forecasting

Distribution of the tweets

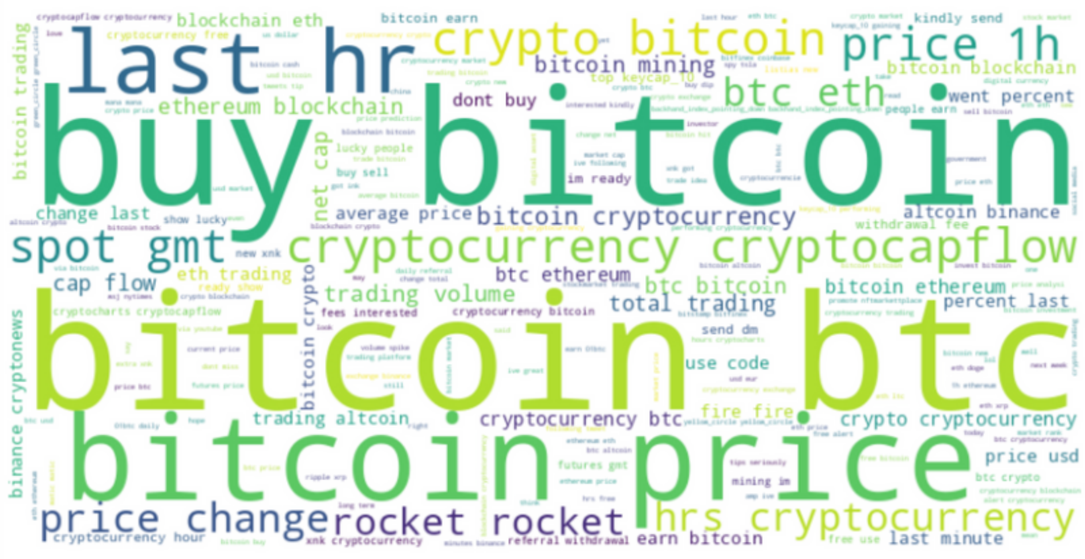


EXPLORATORY DATA ANALYSIS

Preprocessing

- remove links
- remove punctuations
- substitute emojis with theirs description
- remove stopwords

Before



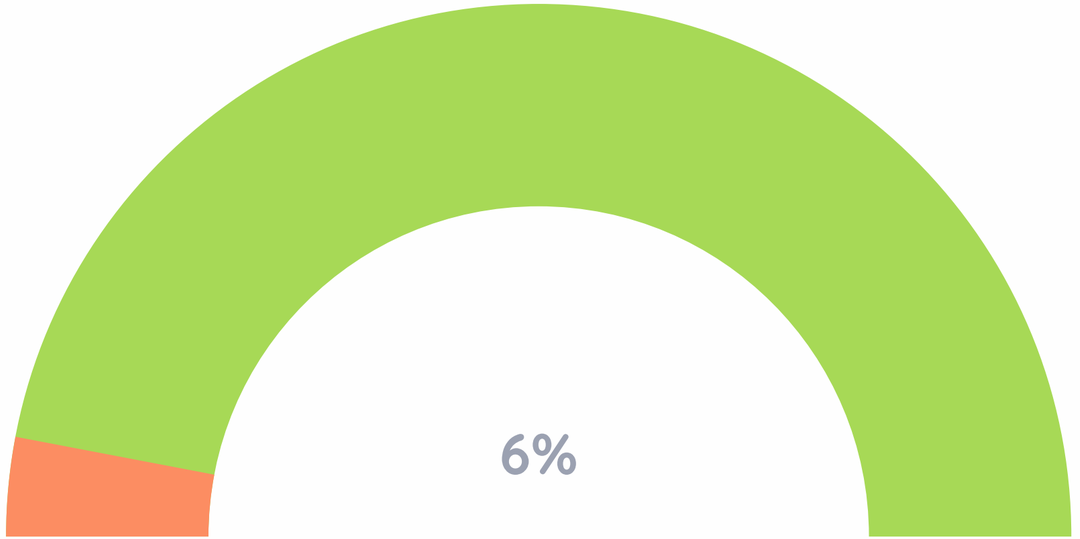
After

Strategy for Avoiding Bot

We removed a total of 52k tweets from the dataset. We eliminated all the tweets that had words like "gits," "gitideas," and "giveaway" in their tweet body.

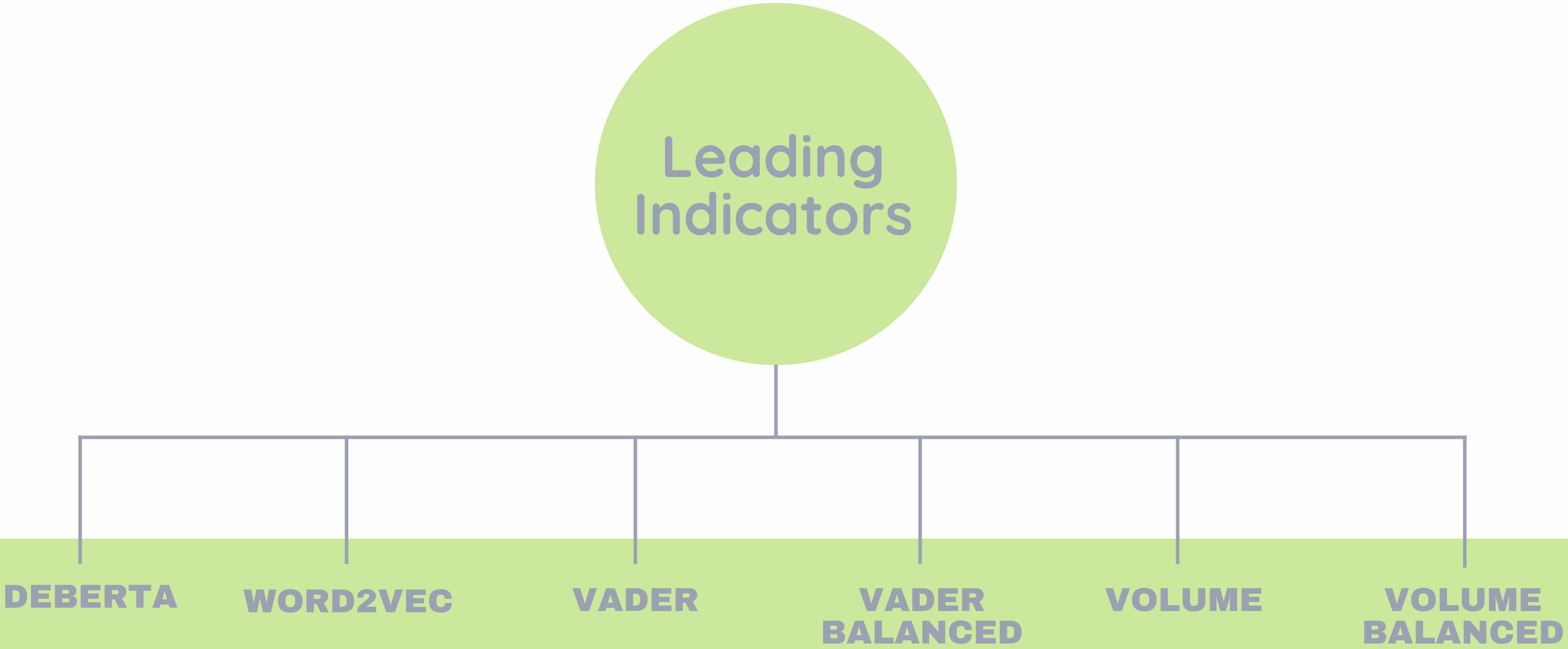
PROFESSIONAL VS THE MAN ON THE STREET

We identified professionals as those who have words like "professor," "researcher," and "economist" in their bio. Additionally, we considered all the verified users as professionals after evaluating the word cloud associated with their user bio.



Bitcoin Price Dataset

We got the Bitcoin USD (BTC-USD) Historical Prices data from Yahoo! Finance for the time range between January 1 2018 to December 31 2022, which is the same time period as the collected tweets.


Leading Indicators
DEBERTA
WORD2VEC
VADER
VADER BALANCED
VOLUME
VOLUME BALANCED

Zero-Shot Sentiment Classification of Bitcoin Tweets using DeBERTa

We only applied the DeBERTa model to two subsets of the dataset: the three days with the highest price peak and the three days with the lowest price peak.

1

Weighted Sentiment Analysis: Incorporating DeBERTa Sentiment Scores with Favorite and Retweet Counts in Bitcoin Tweets

2

Unweighted Sentiment Analysis: Leveraging only DeBERTa Sentiment Scores for Bitcoin Tweets

DeBERTa Index

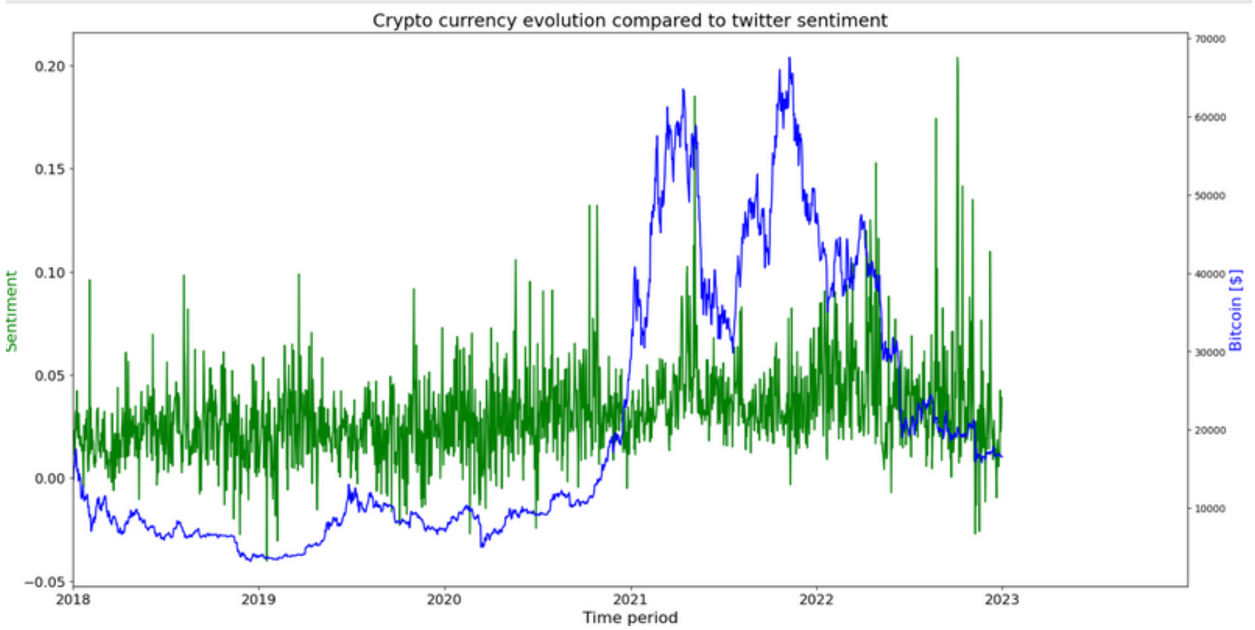
	Highest Price Peak	Lowest Price peak
Professionals	0.67	-0.97
Man on the street	-0.4	0.24

On the side are shown the correlations between the sentiment indicators obtained with DeBERTa and the price of Bitcoin.
We found that, in general, the sentiment of professionals is much more correlated compared to the sentiment of non-professionals.

https://github.com/mmihaltz/word2vec-GoogleNews-vectors

Word2vec

We are using a pretrained Word2Vec model to calculate a score based on positive and negative words associated with Bitcoin tweets.



Positive Words

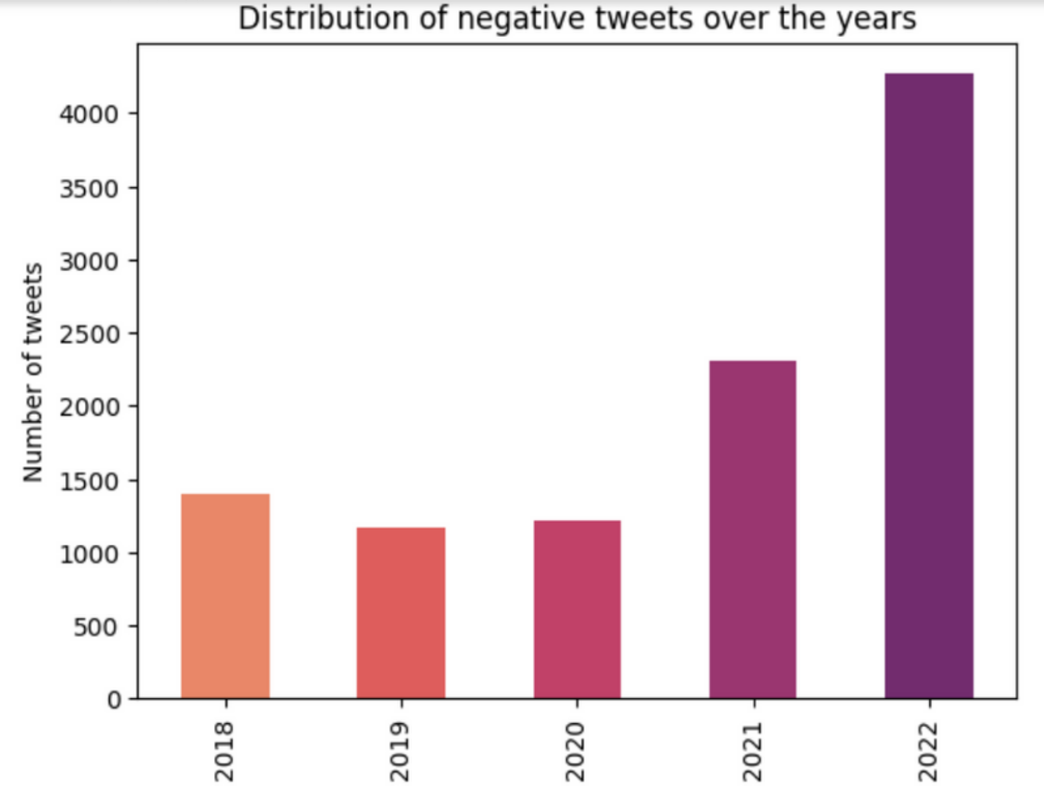
We compute cosine similarity between the word and the positive word list (i.e. bullish, profitable, growth) and add this score to the overall score

Negative words

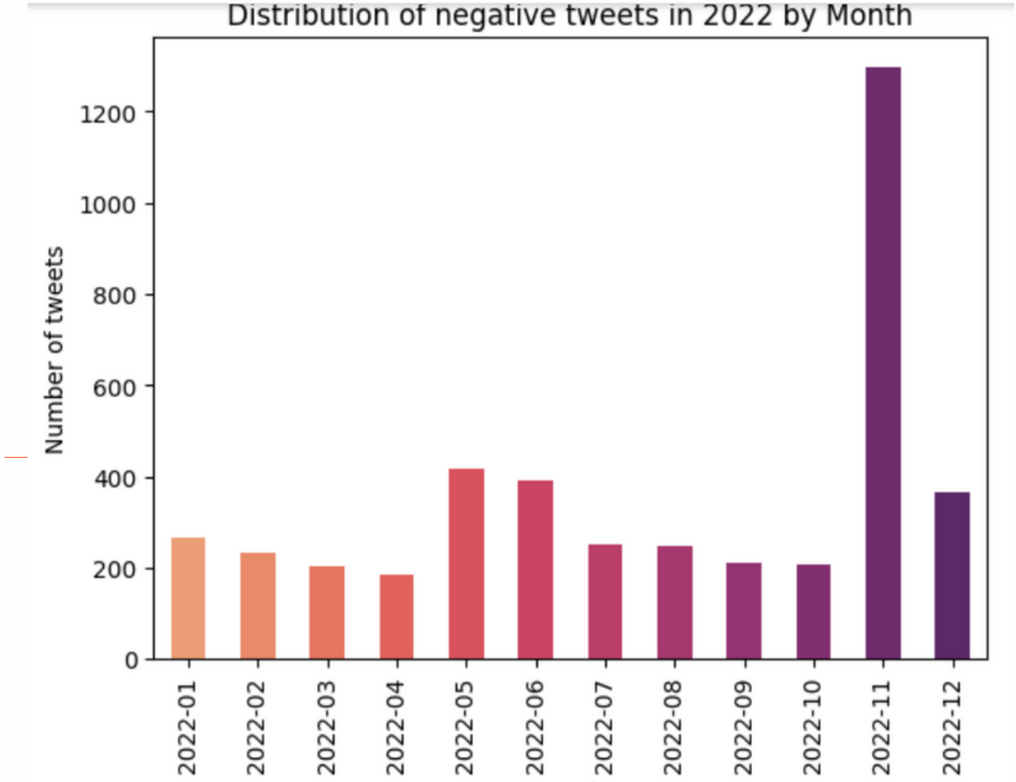
We compute cosine similarity between the word and the negative word list (i.e. bearish, loss) and subtract this score from the overall score.

Word2vec

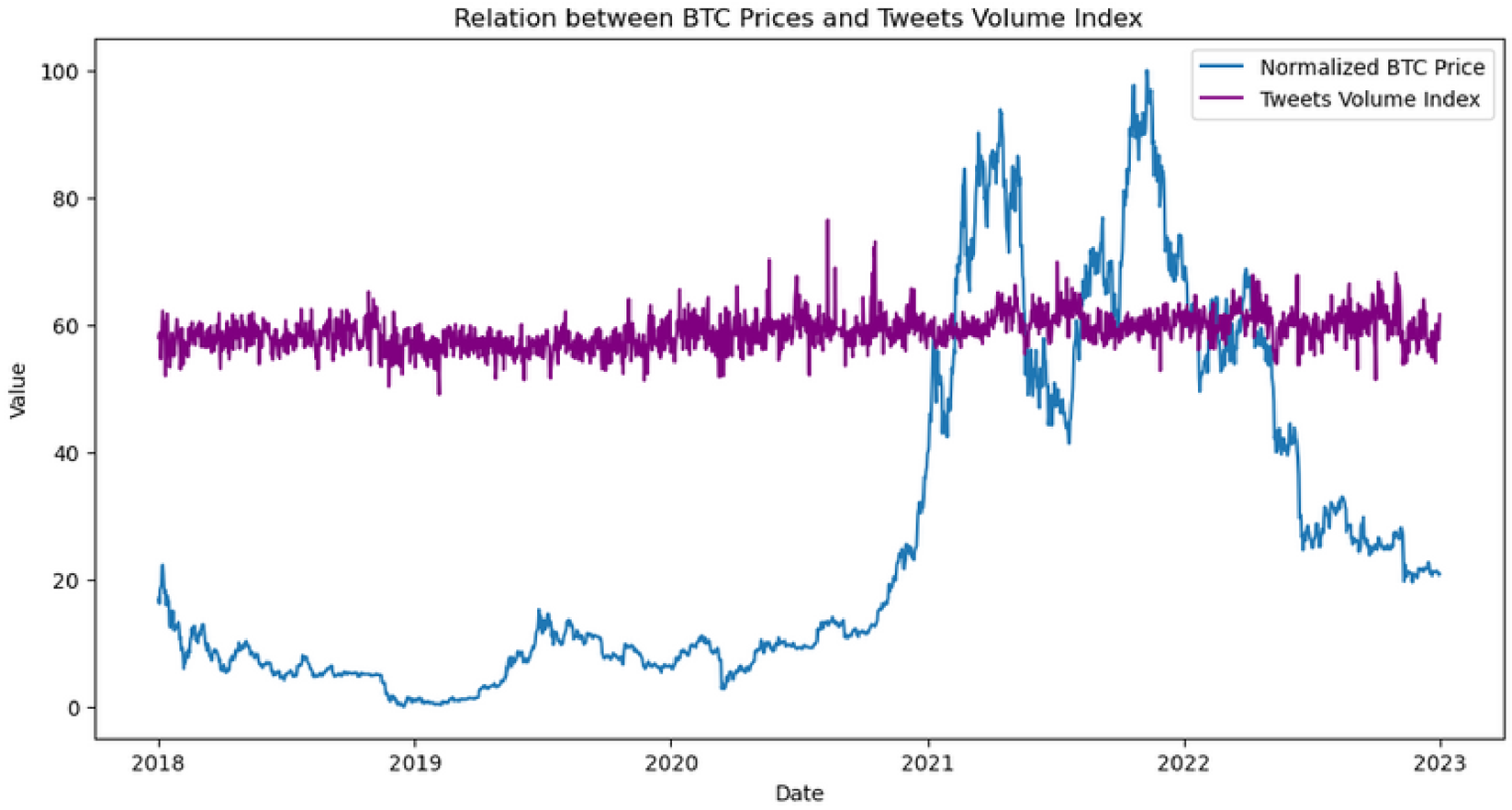
Analyzing Tweets with Negative Sentiment over the Years.



Analyzing Tweets with Negative Sentiment over the Months of 2022.



Vader

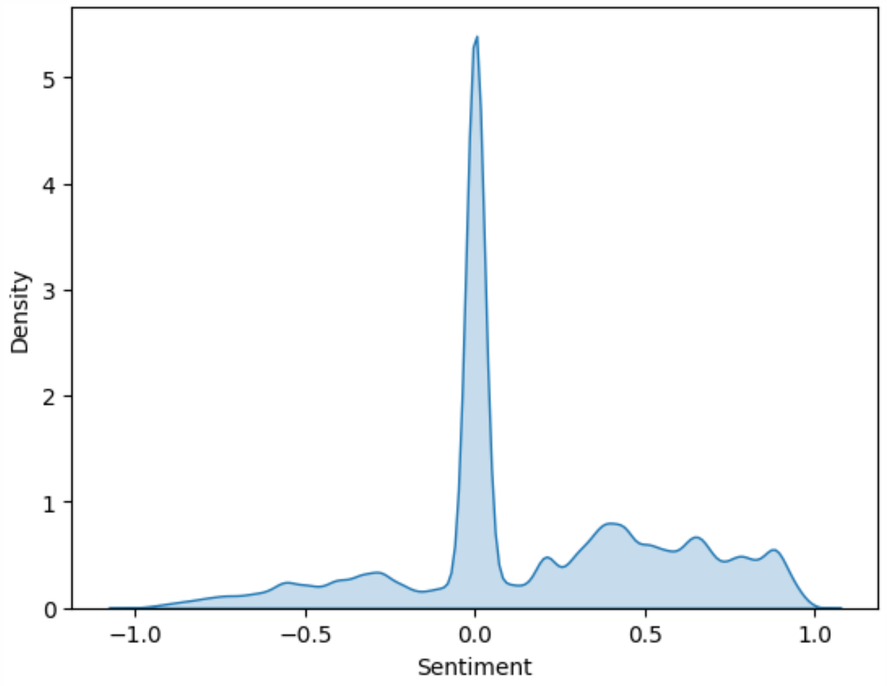


VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media

"Cryptocurrency Price Prediction Using Twitter Sentiment Analysis", N. B. Sahana 2023

Vader: Scores

Density of sentiment distribution

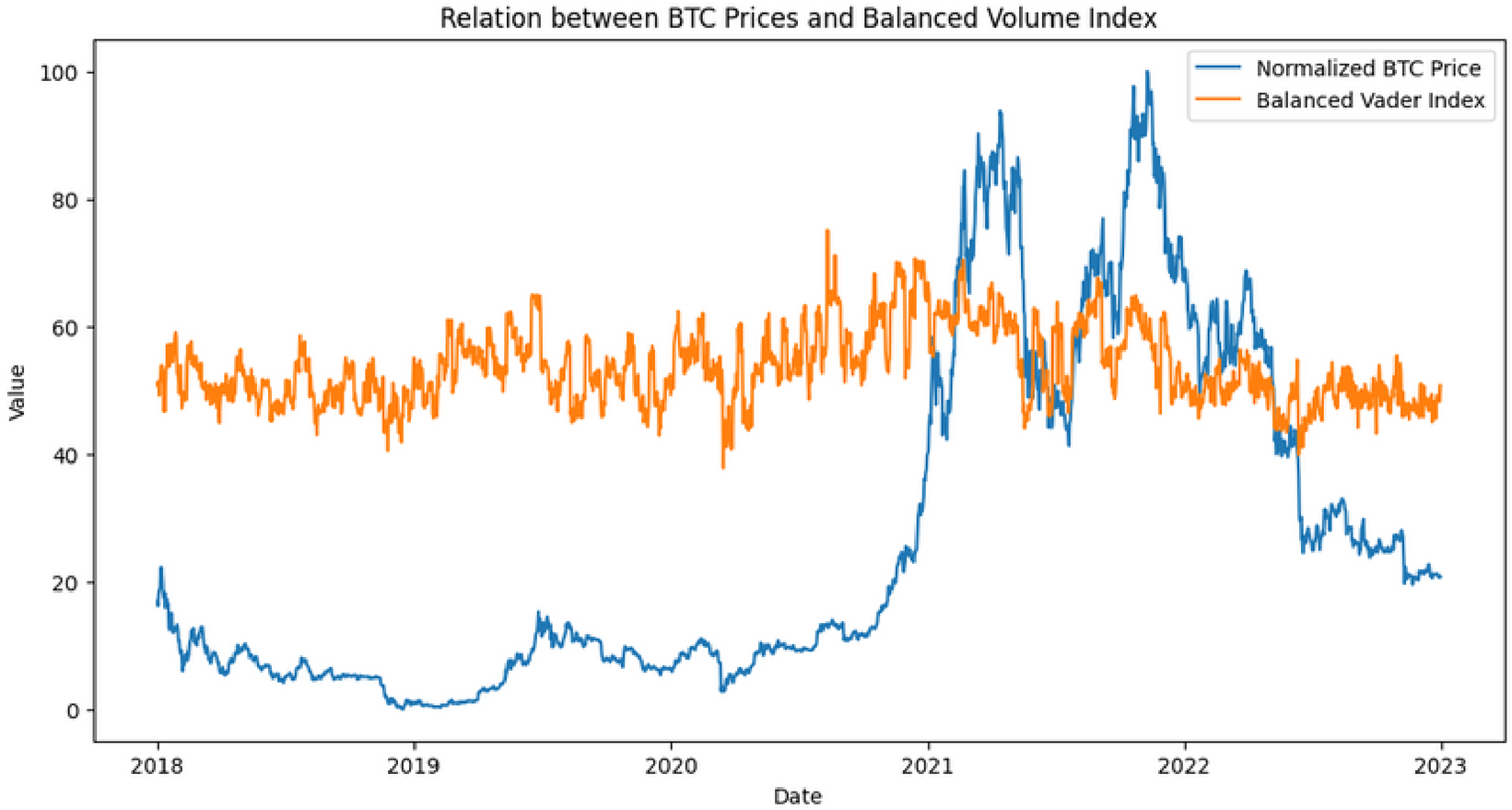


It is observed that the sentiment, according to the VADER score, remains consistently neutral throughout the period.

VADER takes into account negations and contractions (“not good”, “wasn’t good”) punctuation (“good!!!”), capitalized words, text-based ‘emotes’ (for example: “:)”), emotion intensification (very, kind of), acronyms, and scores tweets between -1.0 (negative) and 1.0 (positive).

"Cryptocurrency Price Prediction Using Twitter Sentiment Analysis", N. B. Sahana 2023

Vader Balanced



Obtained from a weighted average between the values from Fear and Greed Index (made by Alternative)[0.2] and the previous Vader index[0.8]

https://alternative.me/crypto/fear-and-greed-index/

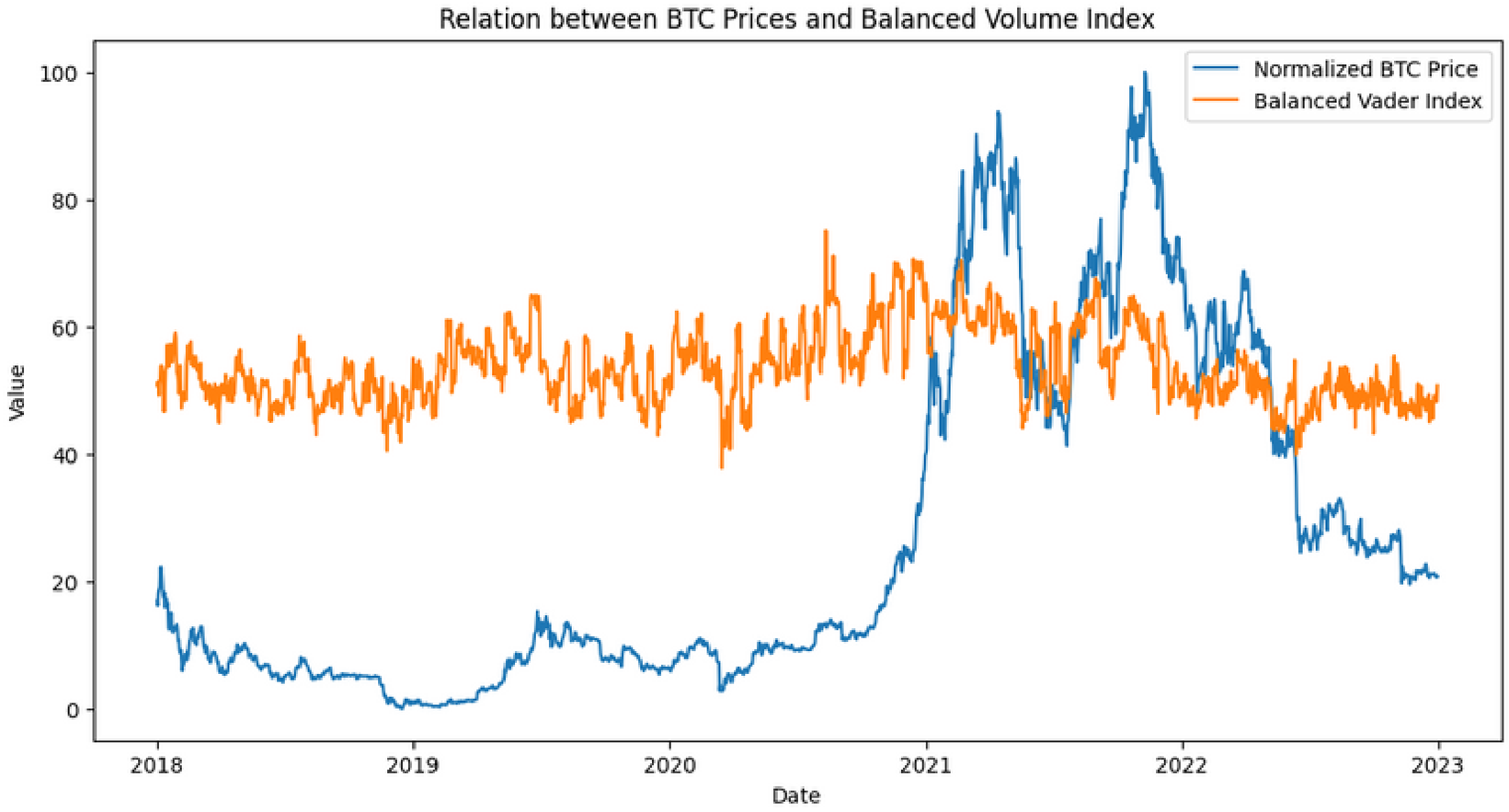
Vader Balanced: Fear And Greed Index

The Fear and Greed Index for cryptocurrencies is a tool that measures current crypto market sentiment based on several indicators:

1. Volatility (25 %)
2. Volume (25%)
3. Social Media (15%)
4. Surveys (15%)
5. Dominance (10%)
6. Google Trends (10%)

https://alternative.me/crypto/fear-and-greed-index/

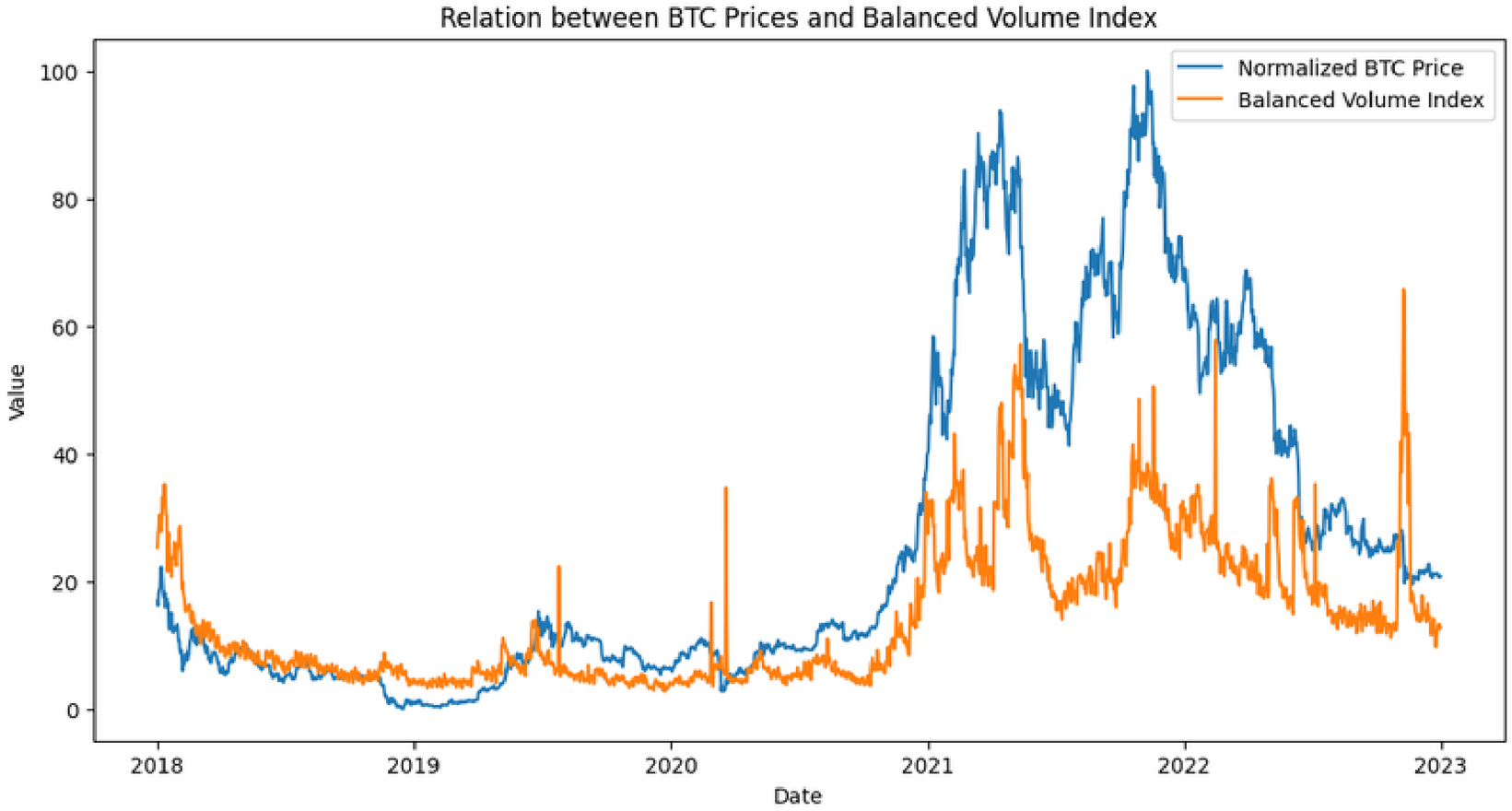
Volume



We constructed this index representing the daily volume of tweets

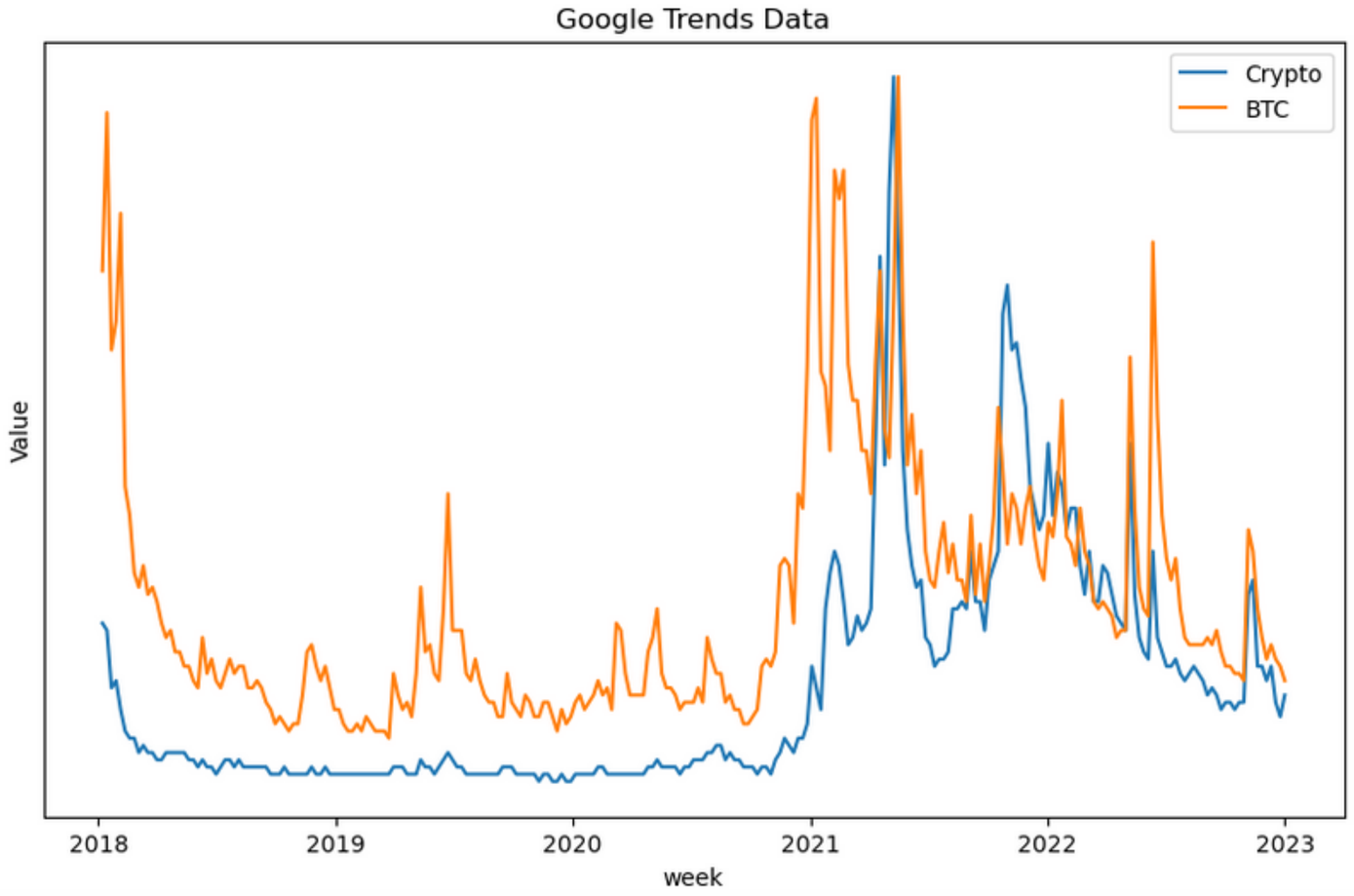
"Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis", J Abraham 2018

Volume Balanced



Obtained from an average between the Google Trend Score (weighted average for the words "Crypto" [0.6] and "Bitcoin" [0.4]) and the previous volume index

Volume Balanced: Google Trend data



we can see that the search trend peaks in 2021 and has the lowest and most constant values in the period 2018-2021. The two curves are highly correlated.

"Do Google Trends Forecast Bitcoins? Stylized Facts and Statistical Evidence", Barrantes 2019

Forecasting Exercise

Long Short Term Memory

XGBoost

ARIMA

Long Short Term Memory

The different indices exhibit a similar trend for the three different metrics considered.
In general, we can consider all the indices to perform well in the forecasting exercise, but word2vec emerges as the most effective performer.

XGBoost

With the XGBoost model, the Vader index turns out to be the best. In particular, the balanced Vader index performs the best in the forecasting exercise.

ARIMA(0,1,0)

Using the arima model, all indexes behave extremely well, and again the word2vec index is the best (although the difference is really small). the mse of all indexes is practically zero.

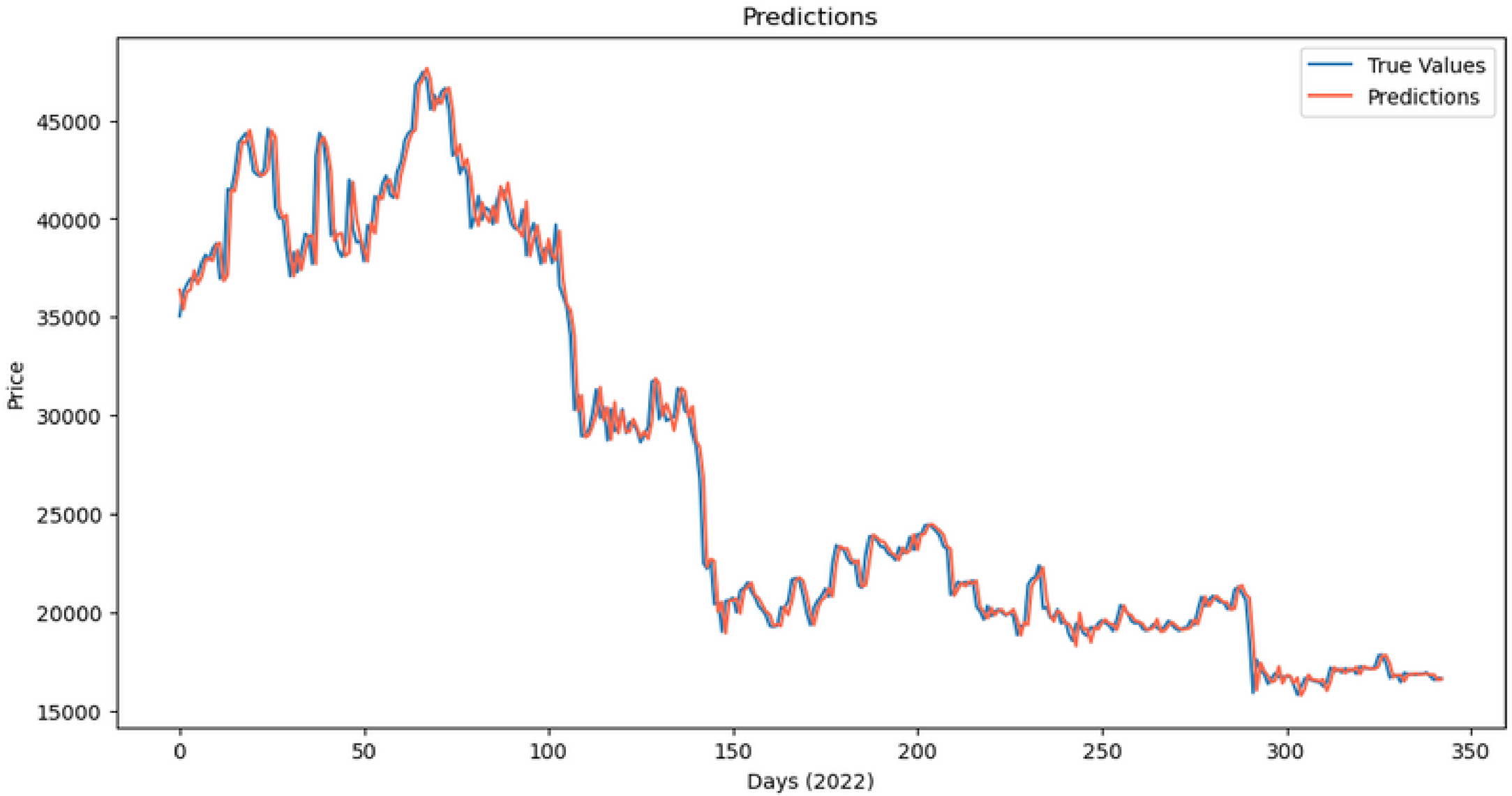
Benchmark: Arima(0,1,0) without indexes

	Benchmark	ARIMA(Word2Vec)
MAE	0.0202	0.0205
MSE	0.0009	0.001
RMSE	0.0314	0.0317

On the side are shown the metrices of our benchmark and the ARIMA computed with the Word2Vec.

We can say that the two models are extremely close in terms of results.

Predictions: ARIMA (Word2Vec)



Thank You

for your attention!