Crypto Streaming

Dennis Shen, Pin-Tsung Huang, Philip Kutlesa
Department of Computer Science
Courant Institute of Mathematical Sciences
New York University
{ts4071, pth254, pk2264}@nyu.edu

Abstract

In this work, we perform real-time Bitcoin (BTC) price prediction based on social media sentiment using Spark Streaming. We process Twitter, Telegram and Binance activity in real-time via socket connections, and engineer input features to a Long Short-Term Memory (LSTM) network to perform BTC price prediction. Our computation was performed on the NYU Peel Cluster.

1 Introduction

In recent years, substantial capital has been invested into the cryptocurrency markets. The average daily trading volume of Bitcoin (BTC) has consistently exceeded 20 billion USD since 2019 (CoinMarketCap, 2021). These markets are also volatile, which presents large financial opportunities and risks for investors. Our motivation for this analytic is to provide actionable insights to investors that are looking to effectively enter and exit the market.

We hypothesize that social media activity is a driver of BTC price. To investigate this, we analyze real-time relations between sentiments on social platforms and the price trends of BTC. When designing an analytical investment tool, we are keen on incorporating the following properties:

Wisdom of the Crowds The collective insights of thousands of users is greater than that of a single expert. We use the sentiment of thousands of social media users to inform price predictions.

Real-time The high volatility of BTC price motivated our decision to use Spark Streaming to develop a real-time processing application. Since the market moves quick, we value timely analytic.

Extensible Analytic Core to our predictive model is that it be extensible in input features. This allows us to experiment with various combinations of features to empirically determine rich feature sets.

1.1 Related Works

We are not the first researchers to investigate the relationships between social media activity and BTC price. In this subsection, we overview related works.

It is proposed evidence that BTC price is motivated by social signals (Coulter, 2022). The work analyzed 4218 articles from 60 international news sources (e.g., Bloomberg) using Latent Dirichlet Allocation (LDA) to model topics that most influence BTC price. They present 18 topics that seemingly impact BTC price within 24 hours of being discussed in large publications during 2018 to 2020.

Another research proposed also uses text analysis techniques and illustrates statistical relations between Twitter activity and BTC price (Mai et al., 2018). This work presents evidence that positive Twitter sentiment corresponds with increases in the future valuation of BTC. Interestingly, they study the impact of individual users on price and conclude that the "silent majority," which are the 95 percent of users that are less active and collectively account for 40 percent of tweets, present the strongest impact on BTC price.

2 Data Sources

In this section, we discuss the data sources used to conduct our analytic. We extract social media data from Twitter and Telegram, and BTC financial data from Binance. All of our data is procured in real-time.

2.1 Twitter Streaming API

Originally, we attempted to use `spark-streaming-twitter` library to

connect to Twitter Streaming API. However, we faced several issues and decided to connect to Twitter Streaming API by shell and open a socket as a source for our streaming code.

In this work, we only query Bitcoin-related tweets. The response format from the Twitter Streaming API is `json`, and there are many columns. We use Spark to extract the columns that we assume have a relationship with BTC price, such as `text`, `followers count`, `created at`, `retweet count`, `reply count`, `like count`, `quote count`. Next, we use Spark NLP to analyze each tweet's sentiment. We classify each tweet as one of three sentiments: `positive`, `negative`, or `na`. After persisting the streaming connection for 2 days, we were able to extract 700K+ records (500 MB).

2.2 Telegram Messages

Telegram Messages are received using the Telethon library[1]. By creating an instance of `TelegramClient` with API key and hash, it is ready to receive messages. The script catches messages by `events.NewMessage(chats=group_name)` which triggers an event handling function whenever a new message coming in and matching the group names we are interested in. When a message comes in, the handler will create a dictionary with three keys–timestamp, group_name, message–and their values. This dictionary then will be append to a temporary list to be further processed. Messages were extracted from 9 Bitcoin-related groups and amounted to 16K+ records in 2 days.

2.3 Binance WebSocket Stream

BTC price and trading data was procured through a persistent WebSocket connection with Binance WebSocket Stream at the kline/candlesticks topic. The stream provides data with the following key properties: `symbol`, `closing price`, `number of trades`, `base volume`, `quote volume`, `daily high`, and `timestamp`. The WebSocket connection pushes updates every second in JSON format. After a persistent 2-day connection, we accumulated 80K+ records containing second-by-second pricing and trading activity data.

[1] https://docs.telethon.dev

3 Methodology

In this work, our methodology consists of the following procedure: (1) procure data via socket connections, (2) process data using operations on DStreams, (3) store results to HDFS, (4) use data as input features to predictive Long Short-Term Memory (LSTM) model, (5) perform price prediction regression task, and (6) visualize results. This procedure is depicted in the design diagram (Figure 1). This is an iterative process which involves continuous feature experimentation and evaluation of analytical results. In this section, we detail the aspects of our methodology.

3.1 Sending over Socket

To send the fetched Tweets, Telegram messages and Binance price data to the Spark program, we use a socket connection with JSON strings encoded in UTF-8. First, we create a socket with the options of `socket.AF_INET`, `socket.SOCK_STREAM`, and bind it to desired `HOSTNAME` and `PORT`. Now the socket listens to incoming connections. If any program (in this case, our Spark program) connects to this socket, the connection is established and remains open as a `conn` object. We will send the encoded JSON string via this connection.

3.2 DStreams and Sentiment Analysis

Connecting to our input data streams via Spark's `socketTextStream` method provides us with DStreams from which we can perform actions and transformations. Data is streamed every second, and DStreams are formed using a 5 minute window. The messages from Twitter and Telegram are sent into an NLP pipeline for further analysis with a pretrained NLP model *ViveknSentimentModel*[2] from John Snow Labs (Kocaman and Talby, 2021) to derive sentiment scores. The resulting data is written as CSV files to HDFS.

3.3 LSTM for Price Prediction

The transformed data that is written to HDFS is loaded into a separate application that executes the training and inference of an LSTM model. The data is used as input features to the neural network that performs price prediction based on the previous 100 minutes of Twitter, Telegram and Binance data.

[2] https://nlp.johnsnowlabs.com/2021/11/22/sentiment_vivekn_en.html

Our motivation to use the LSTM model is its ability to capture statistical relationships across input features and the target variable over long time intervals. This makes it an effective model for handling time-series data. Furthermore, the model is highly extensible. Using the same LSTM model, we are able to utilize various combinations of features to train and evaluate the model.

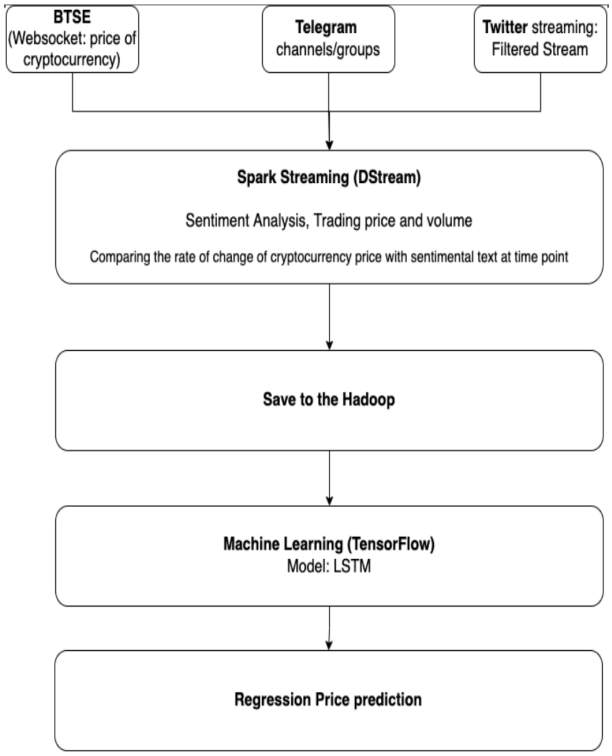

Figure 1: Design Diagram

4 Results

In this section, we present the results of our experiments. This includes the performance of the LSTM model using various feature sets and the relationship between Twitter and Telegram Sentiment with BTC price.

4.1 BTC Price Prediction

We used real-time Twitter and Telegram feature sets as input to the LSTM model to see which datasource was the most informative for BTC price prediction.

Twitter Sentiment The price predictions of the LSTM model using tweet sentiments as input features is presented in Figure 2. As we can see the model's predictions are a couple hundred dollars above the true price. Notably, the model predicts the trajectory of the price well as seen in the strong fit between the predictions and true closing prices.

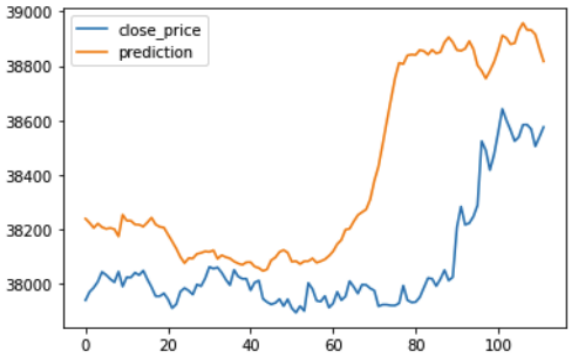

Figure 2: LSTM price predictions using Tweet sentiments as input features

Telegram Sentiment The price predictions of the LSTM model using Telegram message sentiments as input features is presented in Figure 3. It is clear that the LSTM did not have accurate price predictions and did not fit the trajectory of the price well either. We have tried several features including the positive count, the negative count and the amount of messages in a given time frame, but none of the above Telegram information improved the predictions made by the neural network.

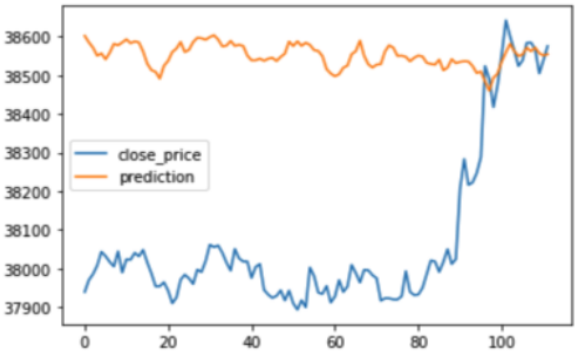

Figure 3: LSTM price predictions using Telegram message sentiments as input features

From these feature set experiments, our results show that Twitter Sentiment is a stronger indicator of BTC price than Telegram sentiment.

4.2 Sentiment vs Price

In Figure 4, we can visualize the relationship between the the aggregate Twitter sentiment regarding Bitcoin and BTC price. In the first half of the trend, we see that the sentiment is net negative and the price of the currency decreases. Likewise, in

the sencond half of the figure, the trend shows a correlation between positive sentiment and price increases.

Figure 4: Tweet Sentiments vs BTC Prices

In contrast, there is a weak relationship between total Telegram sentiment and BTC price. This is depicted in Figure 5.

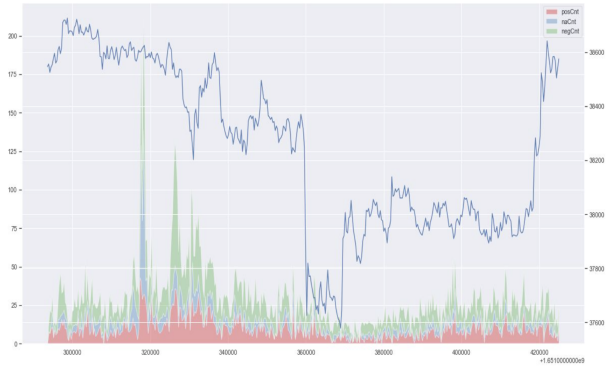

Figure 5: Telegram Sentiment vs BTC Prices

5 Discussion

Our results evaluated Twitter and Telegram sentiments as predictors of BTC price using an LSTM model. Here we discuss our insights and elaborate on our results.

Twitter vs Telegram Our chief result is that Twitter sentiment is a better predictor of BTC price than Telegram sentiment. We stipulate that this is the case because Bitcoin-related tweets reach a larger audience on Twitter than messages do on Telegram. Also, our results show that Twitter sentiment is a good model of the trajectory of Bitcoin price. The key insights from these results is that investors can benefit from tracking Twitter activity to understand whether the price of Bitcoin is likely to increase (positive sentiment) or decrease (negative sentiment).

Streaming Infrastructure Spark Streaming provided us with sufficient infrastructure to process data in real-time across Twitter, Telegram and Binance. Figure 6 shows the streaming performance of processing data from the Binance websocket for more than 2 days. With an average input rate of 0.46 records/sec and more than 80K records processed, it is clear that Spark Streaming provides sufficient capacity to track social media and financial data in real-time.

Limitations For this work, we streamed data from Twitter, Telegram and Binance for 2 days and conducted price predictions for an interval that is more than 8 hours in duration. We do not assume that these results are consistent over time. Procuring data for longer time intervals or using historical data could be viable next steps to validate the consistency of these results. Furthermore, we used one machine learning model to explore the strength of various feature sets. There still remains extensive hyperparameter tuning that could influence the predictive accuracy of the model. It is also likely that different machine learning models respond differently to Twitter and Telegram features than the LSTM did.

6 Conclusion

In this work, we explored the application of Spark Streaming to process social media sentiment in real-time and inform a neural network to perform BTC price predictions. We learned about developing Spark Streaming applications and began to understand the relationship between social media sentiment and BTC price. We found that Twitter is a stronger predictor of BTC price than Telegram. Many variables influence BTC price, and understanding its driving factors remains a continuing field of study.

Correlation does not imply causation, and consequently, we cannot concretely state what does and does not affect the price of cryptocurrencies. However, continued machine learning experimentation, including: feature construction, model selection, and hyperparameter tuning can improve the capacity of our model to more accurately predict BTC price. Perhaps more importantly, exploring new data sources and using more data over longer time intervals could improve the performance of our analytic. Future work includes developing an auto-

mated trading bot that executes trades in real-time based on trading signals. Our findings show that Spark Streaming infrastructure is capable of supporting this development. Practically, this involves further engagement with financial exchange developer APIs to execute trades via software.

7 Acknowledgements

Our research and analytic was made possible by the following:

- Thank Twitter, Telegram and Binance for access to data streams
- Thank Discord community for prompt answers to questions
- Thank NYU HPC for providing platform to develop big data apps
- Thank Keras open-source community for access to neural network models

References

CoinMarketCap. 2021. Bitcoin 24h trade volume from july 1, 2019 to october 28, 2021 (in billion u.s. dollars) [graph]. *Statista*.

Kelly Ann Coulter. 2022. The impact of news media on bitcoin prices: modelling data driven discourses in the crypto-economy with natural language processing. *Royal Society Open Science*, 9(4):220276.

Veysel Kocaman and David Talby. 2021. Spark nlp: Natural language understanding at scale. *Software Impacts*, page 100058.

Feng Mai, Zhe Shan, Qing Bai, Xin (Shane) Wang, and Roger H.L. Chiang. 2018. How does social media impact bitcoin value? a test of the silent majority hypothesis. *Journal of Management Information Systems*, 35(1):19–52.

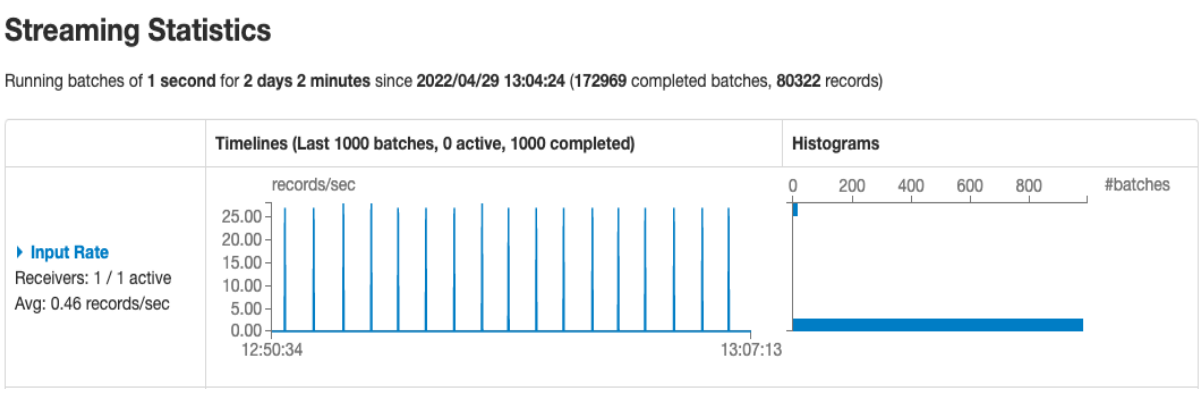

Figure 6: 48 hour data stream reading records every second