Predicting Price Trends in Cryptocurrencies using News Headlines

Daniel Deng, Zain Hemani, and Reid Vender

March 7, 2018

Proposal

Computer Science 4411B
Databases II

Submitted in partial fulfillment of the requirements for the final project

Contents

1 Introduction **2**

2 Motivation **2**

3 Project Details **4**
3.1 Architecture and Environment 4
3.1.1 Available Data 4
3.1.2 Database Structure 5
3.2 Deliverables 7
3.2.1 Features and Finished Product 7
3.2.2 Project Contribution 8

4 Conclusion **8**

1 Introduction

Unlike stocks, mutual funds, commodities and other speculative instruments, cryptocurrencies are highly volatile. This means price predictions are highly variable and can depend on a number of factors, including the public's perceptions of the perceived value of cryptocurrencies. This perceived value is determined by the market, mainly from the media and various news headlines.

Despite few similar applications that use news headlines to predict price trends, many of these applications do not store their collected information in a database. While these applications do many of their calculations in real-time, which may provide useful guidance for investors, storing this information in a database will also allow for regressions and other analyses to be conducted on the datasets to discover long term trends. Our solution involves the use of the NoSQL database MongoDB, R and Python to store and access our data, explore relationships between stocks and cryptocurrencies, and perform MapReduce to analyze keywords which positively or negatively affect cryptocurrency prices.

2 Motivation

When it comes to pricing predictions on cryptocurrencies, we see all sorts of claims: some people say bitcoin will reach $16000, while other says it may skyrocket to $100000 [1]. Unlike stocks, mutual funds, commodities, and other speculative instruments, cryptocurrencies are not regulated and are highly volatile, making such enormous differences in price predictions. This volatility is largely derived from the public's perception of the cryptocurrency and its value. This denotes a clear distinction between the price and value of a cryptocurrency. The price of any cryptocurrency is simply the monetary cost of purchasing it, whereas the value of a cryptocurrency is its perceived benefits and usefulness. The price of a cryptocurrency is not tied to its value but instead, its perceived value. This perceived value is determined in the market, mainly from the media and various news headlines. In many cases, news outlets are the only source of information for many investors in cryptocurrency. This is why correlation with news headlines and the prices of cryptocurrency is an important set of information that can be used to algorithmically and efficiently trade in this ever-growing market.

There is over 20 billion dollars traded each day in various cryptocurrencies around the world. This is a staggering amount which is enormously influenced by news outlets, such as the New York Times, twitter accounts such as John McAfee, and even YouTube videos by large investors. Studies show that tweets by individuals

just as McAfee were reliably correlated with price spikes that sent cryptocurrencies worth pennies — even fractions of a penny — temporarily shooting upwards in value anywhere from 50 to 350 percent. For example, on December 15, McAfee tweeted that a coin called SAFEX constituted "the majority of [his] holdings," [2]. Minutes after McAfee tweeted about SAFEX, the price of SAFEX spiked 92 percent, according to the site CoinMarketCap, from $0.014 USD to a high of $0.027, before settling down for a more gradual rise up to its current price of $0.03 [3]. Another example would be the news headlines related to Korean regulations on cryptocurrency exchanges. Many indexes simply took Korean exchange prices off their representations which caused numerous new outlets, such as the New York Times and Bloomberg to release articles [4, 5]. Although no regulation had actually taken place at the time, due to these articles the prices for cryptocurrencies fell significantly — in some cases by more than 75% — leading to a large market fluctuation [6]. As suggested by the above examples, news headlines provide an important insight into future prices for cryptocurrencies and give investors reliable information when looking to long and short their positions.

There are a few similar applications that use news headlines to predict price trends, however many of these applications do not store their collected information in a database. This is mainly done through natural language processing in tokenizing different segments of a headline or tweet and understanding whether the fluctuation will be positive or negative depending on the statement [7]. These applications do many of their calculations in real-time, at the moment the tweet or news headline is received. This does provide useful guidance for investors, however storing this information in a database, as our application plans to do, will not only allow for real-time inferences but also allow for regressions and other analyses to be conducted on the datasets over a longer period of time. This will help determine long term trends and also differentiate between the extent of effect each news or social media outlet has on prices. Our solution has potential benefits for both individual investors and larger institutional investors. News and social media correlations are already widely used in the industry, but looking at this information from a macro lens will allow investors to correctly ascertain which source is more effective and the potential price fluctuation that may occur from the release. This will certainly provide a useful tool to conduct market analysis.

3 Project Details

3.1 Architecture and Environment

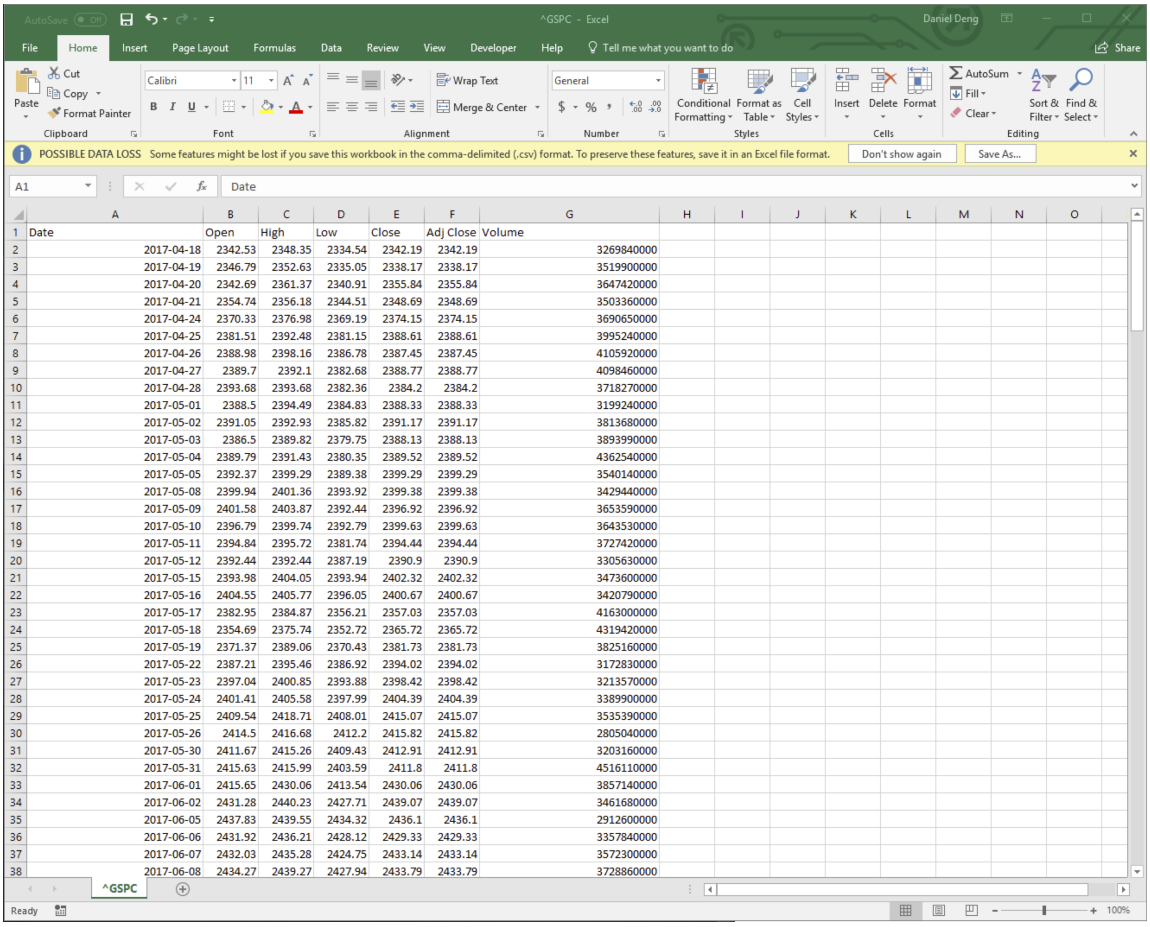
Our project will require datasets from Kaggle and Yahoo Finance. The datasets will come in the form of Excel spreadsheets and they will require cleaning using a python script or built-in functions in Excel. We will be using the NoSQL database MongoDB to store and access our data. Excel and potentially R and Python will be used for our data analysis. Some useful libraries in R include caret, ggplot, dplyr and lattice. PyMongo, NumPy, SciPy, Pandas and Plotly will be essential libraries if we choose to use Python. The hardware that will be used for this project is a standard Windows Laptop with 8Gb of Ram and an Intel i7-7500U processor.

3.1.1 Available Data

The cryptocurrency dataset from Yahoo Finance contains stock opening, high, low and close prices, as well as optional volume, adjusted close and market cap available for dates from 2013 to 2017. "Daily News for Stock Market Prediction" is a dataset from Kaggle which records the top 25 news headlines from /r/worldnews (ranked by Reddit users' votes) from 2008 to 2016 [8]. Together, with over 11,500 stock and cryptocurrency prices and 76,600 news headlines spanning 3,064 days, querying the database could return over 35,000,000 data points.

3.1.2 Database Structure

The figure below shows an example of raw data prices for the S&P (a well-known index that tracks overall market performance). As we can see, there are many different fields but *Date*, *Open* and *Close* will be the most important ones to focus on when we do our analysis.



Date	Open	High	Low	Close	Adj Close	Volume
2017-04-18	2342.53	2348.35	2334.54	2342.19	2342.19	3269840000
2017-04-19	2346.79	2352.63	2335.05	2338.17	2338.17	3519900000
2017-04-20	2342.69	2361.37	2340.91	2355.84	2355.84	3647420000
2017-04-21	2354.74	2356.18	2344.51	2348.69	2348.69	3503360000
2017-04-24	2370.33	2376.98	2369.19	2374.15	2374.15	3690650000
2017-04-25	2381.51	2392.48	2381.15	2388.61	2388.61	3995240000
2017-04-26	2388.98	2398.16	2386.78	2387.45	2387.45	4105920000
2017-04-27	2389.7	2392.1	2382.68	2388.77	2388.77	4098460000
2017-04-28	2393.68	2393.68	2382.36	2384.2	2384.2	3718270000
2017-05-01	2388.5	2394.49	2384.83	2388.33	2388.33	3199240000
2017-05-02	2391.05	2392.93	2385.82	2391.17	2391.17	3813680000
2017-05-03	2386.5	2389.82	2379.75	2388.13	2388.13	3893990000
2017-05-04	2389.79	2391.43	2380.35	2389.52	2389.52	4362540000
2017-05-05	2392.37	2399.29	2389.38	2399.29	2399.29	3540140000
2017-05-08	2399.94	2401.36	2393.92	2399.38	2399.38	3429440000
2017-05-09	2401.58	2403.87	2392.44	2396.92	2396.92	3653590000
2017-05-10	2396.79	2399.74	2392.79	2399.63	2399.63	3643530000
2017-05-11	2394.84	2395.72	2381.74	2394.44	2394.44	3727420000
2017-05-12	2392.44	2392.44	2387.19	2390.9	2390.9	3305630000
2017-05-15	2393.98	2404.05	2393.94	2402.32	2402.32	3473600000
2017-05-16	2404.55	2405.77	2396.05	2400.67	2400.67	3420790000
2017-05-17	2382.95	2384.87	2356.21	2357.03	2357.03	4163000000
2017-05-18	2354.69	2375.74	2352.72	2365.72	2365.72	4319420000
2017-05-19	2371.37	2389.06	2370.43	2381.73	2381.73	3825160000
2017-05-22	2387.21	2395.46	2386.92	2394.02	2394.02	3172830000
2017-05-23	2397.04	2400.85	2393.88	2398.42	2398.42	3213570000
2017-05-24	2401.41	2405.58	2397.99	2404.39	2404.39	3389900000
2017-05-25	2409.54	2418.71	2408.01	2415.07	2415.07	3535390000
2017-05-26	2414.5	2416.68	2412.2	2415.82	2415.82	2805040000
2017-05-30	2411.67	2415.26	2409.43	2412.91	2412.91	3203160000
2017-05-31	2415.63	2415.99	2403.59	2411.8	2411.8	4516110000
2017-06-01	2415.65	2430.06	2413.54	2430.06	2430.06	3857140000
2017-06-02	2431.28	2440.23	2427.71	2439.07	2439.07	3461680000
2017-06-05	2437.83	2439.55	2434.32	2436.1	2436.1	2912600000
2017-06-06	2431.92	2436.21	2428.12	2429.33	2429.33	3357840000
2017-06-07	2432.03	2435.28	2424.75	2433.14	2433.14	3572300000
2017-06-08	2434.27	2439.27	2427.94	2433.79	2433.79	3728860000

In order to store the previously described available data, we will first create a single database. Next, we will populate a single collection with an embedded structure:

```
{
  date: <date>,
  headlines: <array>,
  prices: [
    {
      name: <string>,
      open: <double>,
      high: <double>,
      low: <double>,
      close: <double>,
      volume: <long>,
      adj_close: <double>,
      market_cap: <long>
    },
    ...
  ]
}
```

3.2 Deliverables

3.2.1 Features and Finished Product

The project will produce a report on our findings. Details of this report are outlined in the deliverables below. Our project will consist of the following:

Deliverable	**Details**
Deliverable 1	Find appropriate datasets and clean the data so that it is consistent and usable.
Deliverable 2	Import the dataset into a NoSQL database and perform relevant queries to extract data for analysis.
Deliverable 3	Perform MapReduce on the world news data to extract and count keywords present in the news headlines.
Deliverable 4	Analyze data for: • Correlations in stock and cryptocurrency prices over time • Keywords in world news headlines which affect stock and cryptocurrency prices • The stock market is open between 9:30-4pm whereas the cryptocurrency market is open 24/7; the time when the cryptocurrency market is open but the stock market is not is an area of interest
Deliverable 5	(Optional) Build a predictive model using a decision tree like random forest or xgb to forecast future direction/prices based on keywords in news headlines.
Deliverable 6	(Optional) Build a front-end to showcase our findings interactively for the user.
Deliverable 7	Write a report on the findings of our data analysis and its relevance to the industry, and any future improvements that could be made.

The overarching goal of our project is to determine the relationships, if any, between cryptocurrency prices and a benchmark index like the S&P, and keywords present in news headlines. Finding a correlation between these datasets can open the doors to a myriad of new trading strategies. Not only are we striving for monetary gain with this project, we are hoping our research will help the average investor improve their investment decisions. Some personal achievements undertaking such a project include getting hands-on experience with setting up and accessing data from a NoSQL database and learning basic data analytics skills.

3.2.2 Project Contribution

As indicated, some of these deliverables are optional but this project is of great interest to us. As investors ourselves, this project can help us explore trading strategies we have never tried before while improving our technical skills. Our unique, database-oriented solution will also allow for regressions and other analyses to be conducted on the datasets to discover long term trends which we hope will be useful guidance for investors. We will be continuing to work on this assignment in the future and seeing it through to the end.

4 Conclusion

Enormous differences in price predictions are, in large part, due to the high volatility of cryptocurrencies. This volatility is largely derived from the public's perception of the cryptocurrency and its value. This distinction between the price and value of a cryptocurrency may be due to news headlines which shape the public's perception. In order to explore this, we plan to design a NoSQL database using MongoDB to store and access our data, R and Python to analyze relationships between stocks and cryptocurrencies, and perform MapReduce to analyze keywords which positively or negatively affect cryptocurrency prices. This is a different approach from similar applications that use news headlines to predict price trends in real-time without the use of databases. This useful tool to conduct market analysis will help us explore trading strategies we have never tried before while improving our technical skills and has potential benefits for both individual investors and larger institutional investors.

References

[1] A. Morris, "Bitcoin prediction 2018: Here's where investors see bitcoin going THIS YEAR," February, 2018. [Online]. Available: https://www.express.co.uk/finance/city/913707/Bitcoin-prediction-2018-price-crash-bubble-Ripple-ethereum-cryptocurrency-blockchain. [Accessed Feb. 23, 2018].

[2] J. McAfee, December, 2017. [Online]. Available: https://twitter.com/officialmcafee

[3] CoinMarketCap, "Safe Exchange Coin," December, 2017. [Online]. Available: https://coinmarketcap.com/currencies/safe-exchange-coin/

[4] K. Cho, "Why the Cryptocurrency World Is Watching South Korea," February, 2018. [Online]. Available: https://www.bloomberg.com/news/articles/2018-02-04/why-the-cryptocurrency-world-is-watching-south-korea-quicktake. [Accessed Feb. 23, 2018].

[5] REUTERS, "South Korea Keeps Investors Guessing on Cryptocurrency Regulation," February, 2018. [Online]. Available: https://www.nytimes.com/reuters/2018/02/27/business/27reuters-crypto-currencies-southkorea.html. [Accessed Feb. 27, 2018].

[6] REUTERS, "South Korea Regulator Flags Better Deal for Cryptocurrency Industry," February, 2018. [Online]. Available: https://www.nytimes.com/reuters/2018/02/20/business/20reuters-southkorea-cryptocurrencies.html. [Accessed Feb. 23, 2018].

[7] E. R. Connor Lamon, Eric Nielsen, "Cryptocurrency Price Prediction Using News and Social Media Sentiment." [Online]. Available: http://cs229.stanford.edu/proj2017/final-posters/5138197.pdf. [Accessed Feb. 23, 2018].

[8] Aaron7sun, "Daily news for stock market prediction," March, 2016. [Online]. Available: https://www.kaggle.com/aaron7sun/stocknews. [Accessed Feb. 25, 2018].