Bitcoin Price Prediction Using Social Media

Sree Lakshmi Addepalli
NYU Courant
New York, United States

Divya Juneja
NYU Courant
New York, United States

Sree Gowri Addepalli
NYU Courant
New York, United States

Abstract—

The project aims to build a Bitcoin price prediction system by leveraging the data in social media. The application uses tweets and others sources of social media ie. Reddit to correlate that a change in a sentiment about Bitcoins in the social media does impact the bitcoin price and this model generated is used to predict bitcoin price.

I. INTRODUCTION

The social impact of crypto currencies is growing fast along with social media posts like tweets and reddit. The financial markets show to be in a relationship between media sentiment value and the prices of cryptocurrency coins. Sentiment analysis on available online media can inform predictions on whether a coin's price will go up or down. Specifically, in this project we try to answer questions whether sentiment analysis on reddit comments and social media posts like twitter produce accurate predictions about the future price fluctuations of bitcoin.

II. MOTIVATION

(Write a paragraph describing why you think this application is important.)

III. RELATED WORK

In [1] they looked into how altcoins can be used to predict from social media. In order to do this, they collected a dataset containing prices and the social media activity of 181 altcoins in the form of 426,520 tweets over a timeframe of 71 days. The containing public mood were then estimated using sentiment analysis. To predict altcoin returns, they carried out linear regression analyses based on 45 days of data. They showed that short-term returns can be predicted from activity and sentiments on Twitter.

In [2] the paper uses time-series analysis to study the relationship between Bitcoin prices and fundamental economic variables, technological factors and measurements of collective mood derived from Twitter feeds. Sentiment analysis has been performed on a daily basis through the utilization of a state-of-the art machine learning algorithm, namely Support Vector Machines (SVMs). A series of short-run regressions showed that the Twitter sentiment ratio is positively correlated with Bit coin prices.

In [3] the paper showed the Interaction between media sentiment and the Bitcoin price. Here the database of relative news articles as well as blog posts has been collected for the purpose of this research. Hence, each article had been given a sentiment score depending on the negative and positive words used in the article. It has identified that interaction between media sentiment and the Bitcoin price exists, and that there is a tendency for investors to overreact on news in a short period of time.

In [4] the paper used education levels of educations or users who are interested in crypto currencies using 7 readability techniques. Reddit comments were used as part of profiling. This helped new investors and users get profile information about the users interested in cryptocurrencies.

In [5] the paper analyzes the ability of news and social media data to predict price fluctuations for three cryptocurrencies: bitcoin, litecoin and ethereum. Traditional supervised learning algorithms were utilized for text based sentiment classification, but with a twist. Daily news and social media data were labeled based on actual price changes one day in the future for each coin, rather than on positive or negative sentiment. By taking this approach, the model is able to directly predict price fluctuations instead of needing to first predict sentiment. The final version of the model was able to correctly predict, on average, the days with the largest percent increases and percent decreases in price for bitcoin and ethereum over the 67 days encompassing the test set.

In [6], they attempt to make price predictions using news data in addition to social media data. Secondly, they predict price fluctuations for Ethereum and Litecoin in addition to bitcoin. Third, they label our data based on actual price changes rather than traditional text sentiment. Finally, and most importantly, they analyze predicted price fluctuations by size (i.e., percent change); this is a valuable data point to know given the extreme daily volatility of current cryptocurrency markets.

In [7] they collected BTC/USD historical exchange rates. CoinDesk API was used to collect OHLC (Open High Low Close prices) data and collected Twitter's messages in real time. Researchers have used open source python framework which works with Twitter's public API, called Tweepy. To collect relevant tweets these synonyms were used: Bitcoin, BTC, XBT, satoshi. Total of 2,271,815 Bitcoin related tweets were gathered from the Twitter API. Noise reduction in gathered data. All non-alphabetic symbols were removed from tweets. Also, non-unique, suspicious intent or bot generated tweets were discarded. Students 10 have analyzed 100 thousand tweets manually to find suspicious bot accounts, hashtags, words or sentences. Sentiment analysis using VADER open source software. It evaluates every tweet with score from -1 (very negative) to 1 (very positive), which is saved as an extra column in dataset. VADER creators claim that it outperforms most of popular AI sentiment tools when analyzing social media short length messages. Aggregating sentiment scores. Tweets are grouped in a selected timeframe and their sentiment scores are averaged, to get average sentiment for that timeframe. Price prediction is tested with different timeframes7 and shifts forward to find the best fit.

In [8] the authors use time-series analysis to study the relationship between Bitcoin prices and fundamental economic variables, technological factors and measurements of collective mood derived from Twitter feeds. Sentiment analysis has been performed daily through the utilization of a state of-the-art machine learning algorithm, namely Support Vector Machines (SVMs). A series of short-run regressions shows that the Twitter sentiment ratio is positively correlated with Bitcoin prices. The short-run analysis also reveals that the number of Wikipedia search queries and the hash rate (measuring the mining difficulty) have a positive effect on the price of Bitcoins. On the contrary, the value of Bitcoins is negatively affected by the exchange rate between the USD and the euro. A vector error-correction model is used to investigate the existence of long-term relationships between cointegrated variables. This kind of long-run analysis reveals that the Bitcoin price is positively associated with the number of Bitcoins in circulation (representing the total stock of money supply) and negatively associated with the Standard and Poor‟s 500 stock market index.

In [9] the authors analyze correlations and causalities between bitcoin market indicators and Twitter posts containing emotional signals on Bitcoin. Within a timeframe of 104 days, about 160,000 Twitter posts containing "bitcoin" and a positive, negative or uncertainty related term were collected and further analyzed. For instance, the terms "happy", "love", "fun", "good", "bad", "sad" and "unhappy" represent positive and negative emotional signals, while "hope", "fear" and "worry" are considered as indicators of uncertainty. The static (daily) Pearson correlation results show a significant positive correlation between emotional tweets and the close price, trading volume and intraday price spread of Bitcoin. However, a dynamic Granger causality analysis does not confirm a statistically significant effect of emotional Tweets on Bitcoin market values. To the contrary, the analyzed data shows that a higher Bitcoin trading volume Granger causes more signals of uncertainty within a 24 to 72-hour timeframe. This result leads to the interpretation that emotional sentiments rather mirror the market than that they make it predictable. Finally, the conclusion of this paper is that the microblogging platform Twitter is Bitcoin's virtual trading floor, emotionally reflecting its trading dynamics.

In [10] the authors ascertain with what accuracy the direction of Bitcoin price in USD can be predicted. The price data is sourced from the Bitcoin Price Index. The task is achieved with varying degrees of success through the implementation of a Bayesian optimized recurrent neural network (RNN) and a Long Short-Term Memory (LSTM) network. The LSTM achieves the highest classification accuracy of 52% and a RMSE of 8%. The popular ARIMA model for time series forecasting is implemented as a comparison to the deep learning models. As expected, the non-linear deep learning methods outperform the ARIMA forecast which performs poorly. Finally, both deep learning models are benchmarked on both a GPU and a CPU with the training time on the GPU outperforming the CPU implementation by 67.7%.

In [11], the authors present a method for predicting changes in Bitcoin and Ethereum prices utilizing Twitter data and Google Trends data. Bitcoin and Ethereum, the two largest cryptocurrencies in terms of market capitalization represent over $160 billion dollars in combined value. However, both Bitcoin and Ethereum have experienced significant price swings on both daily and long-term valuations. Twitter is increasingly used as a news source influencing purchase decision by informing users of the currency and its increasing popularity. As a result, quickly understanding the impact of tweets on price direction can provide a purchasing and selling advantage to a cryptocurrency user or a trader. By analyzing tweets, they found that tweet volume, rather than tweet sentiment (which is invariably overall positive regardless of price direction), is a predictor of price direction. By utilizing a linear model that takes as input tweets and Google Trends data, they were able to accurately predict the direction of price changes. By utilizing this model, a person can make better informed purchase and selling decisions related to Bitcoin and Ethereum.

In [12] they analyzed 2.27 million Bitcoin-related tweets for sentiment fluctuations that could indicate a price change soon. This is done by a naive method of solely attributing rise or fall based on the severity of aggregated Twitter sentiment change over periods ranging between 5 minutes and 4 hours, and then shifting these predictions forward in time 1, 2, 3 or 4 time periods to indicate the corresponding BTC interval time. The prediction model evaluation showed that aggregating tweet sentiments over a 30 min period with 4 shifts forward, and a sentiment change threshold of 2.2%, yielded a 79% accuracy.

In [13], the author tried to predict the future price of bitcoin in a shorter period. He implemented a lot of trading indicators and technical analysis techniques used in financial stocks followed by machine learning techniques to learn from these indicators and predict the future price of bitcoin. He used convolutional Neural Network (CNN) with technical indicators to obtain unto 90% accuracy.

In [14], this the authors first, we gathered the data relevant to Bitcoin for the purpose of the experiment. More specifically, Bitcoin-related posts on the online forum, daily Bitcoin transaction counts, and its price were gathered. They also extracted and rated significant keywords from the data gathered on the online forum. Then, they selected the data of higher score ratings to generate the prediction model based on deep learning and used the model to predict the fluctuation in the Bitcoin price and transaction count.

In [15], the authors try to prove whether Twitter data relating to cryptocurrencies can be utilized to develop advantageous crypto coin trading strategies. By way of supervised machine learning techniques, the team will outline several machine learning pipelines with the objective of identifying cryptocurrency market movement. The prominent alternative currency examined in this paper is Bitcoin (BTC). Our approach to cleaning data and applying supervised learning algorithms such as logistic regression, Naive Bayes, and support vector machines leads to a final hour-to-hour and day-to-day prediction accuracy exceeding 90%. In order to achieve this result, rigorous error analysis is employed in order to ensure that accurate inputs are utilized at each step of the model. This analysis yields a 25% accuracy increase on average.

IV. Application Design

3 datasets are used as big data and stored in HDFS for cleaning and profiling using Spark. MlLib is used for running Machine Learning algorithms and models like LSTM to analyze relationship of rise and fall in bitcoin price with sentiments of users and also getting future bitcoin predictions. Predictions are sent as notifications on Twilio. Visualization is done using Tableau.

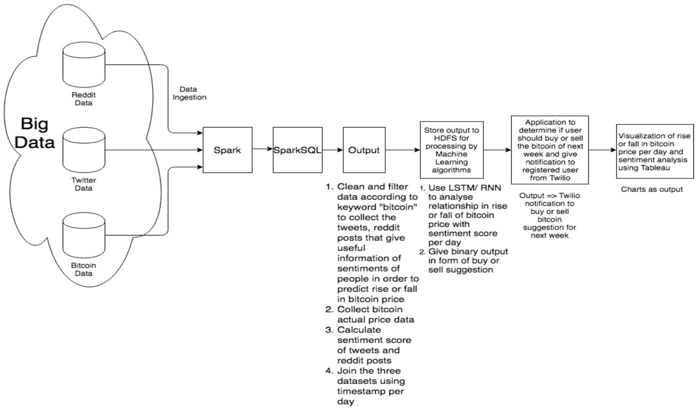
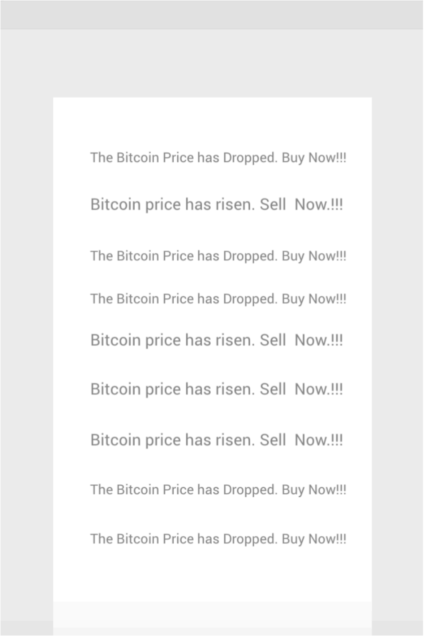

Big Data
Reddit Data
Twitter Data
Bitcoin Data
Data Ingestion
Spark
SparkSQL
Output
Store output to HDFS for processing by Machine Learning algorithms
Application to determine if user should buy or sell the bitcoin of next week and give notification to registered user from Twilio
Visualization of rise or fall in bitcoin price per day and sentiment analysis using Tableau
1. Clean and filter data according to keyword "bitcoin" to collect the tweets, reddit posts that give useful information of sentiments of people in order to predict rise or fall in bitcoin price
2. Collect bitcoin actual price data
3. Calculate sentiment score of tweets and reddit posts
4. Join the three datasets using timestamp per day
1. Use LSTM/ RNN to analyse relationship in rise or fall of bitcoin price with sentiment score per day
2. Give binary output in form of buy or sell suggestion
Output => Twilio notification to buy or sell bitcoin suggestion for next week
Charts as output

Fig 1: Project Design Diagram

UI is just the notification sent to user on Twilio regarding future bitcoin price prediction.



The Bitcoin Price has Dropped. Buy Now!!!

Bitcoin price has risen. Sell Now.!!!

The Bitcoin Price has Dropped. Buy Now!!!

The Bitcoin Price has Dropped. Buy Now!!!

Bitcoin price has risen. Sell Now.!!!

Bitcoin price has risen. Sell Now.!!!

Bitcoin price has risen. Sell Now.!!!

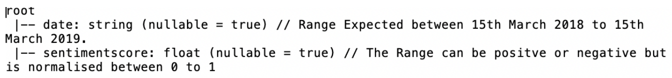
The Bitcoin Price has Dropped. Buy Now!!!

The Bitcoin Price has Dropped. Buy Now!!!

Fig 2: UI for sending notifications on Twilio

V. Datasets

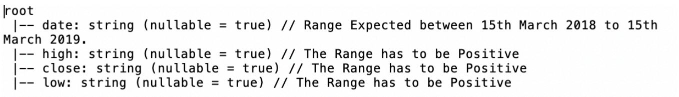
1 year Reddit dataset is fetched from pushshift.io using subreddit 'bitcoin' and collected from 15-03-2018 till 15-03-2019. Reddit blogs describe feelings and sentiments of users related to prices fluctuations of bitcoin.



```
root
 |-- date: string (nullable = true) // Range Expected between 15th March 2018 to 15th
March 2019.
 |-- sentimentscore: float (nullable = true) // The Range can be positve or negative but
is normalised between 0 to 1
```

Fig 3: Schema of Reddit dataset

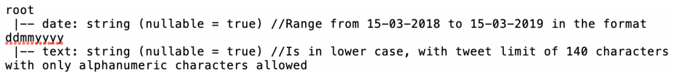
Second dataset is actual bitcoin price using coinmarketcap, (https://coinmarketcap.com/currencies/bitcoin/historical-data/). Opening, closing, high, low, volume, market cap data is scraped from 15-03-2018 till 15-03-2019. Bitcoin price is scraped in USD.



```
root
 |-- date: string (nullable = true) // Range Expected between 15th March 2018 to 15th
March 2019.
 |-- high: string (nullable = true) // The Range has to be Positive
 |-- close: string (nullable = true) // The Range has to be Positive
 |-- low: string (nullable = true) // The Range has to be Positive
```

Fig 4: Schema of Bitcoin actual price dataset

Third dataset, i.e., Twitter dataset is fetched using scraping of data of past 1 year. Bitcoin hashtags are scraped and tweets are used to get sentiment score and analyze the users sentiments.



```
root
 |-- date: string (nullable = true) //Range from 15-03-2018 to 15-03-2019 in the format
ddmmyyyy
 |-- text: string (nullable = true) //Is in lower case, with tweet limit of 140 characters
with only alphanumeric characters allowed
```

Fig 5: Schema of Twitter dataset

VI. Remediation

(Describe the actuation/remediation response to the actionable insight. Is it automated (or possible to automate), or does it require human intervention.)

VII. Experiments

(In this section, you can describe: Your experimental setup, problems with: data, performance, tools, platforms, etc. Discuss your experiments, describe what you learned. Discuss limitations of the application. Discuss what you would do to expand it given time - how would you improve it, etc.)

VIII. Conclusion

(One paragraph about the value, results, usefulness of your application.)

IX. Future Work

(Discuss possible future work for extending this project.)

Acknowledgment

(This section is optional. It can be used to thank the people/companies/organizations who have made data available to you, for example. You can list any HPC people who were particularly helpful, if you used the NYU HPC. List Amazon if you used an Amazon voucher.)

References

(Add references for all of the papers, texts, and data sources that you refer to in your paper. You may have websites to reference, the Spark book, the Hadoop book, etc. A reference is added below as an example.)

1. Christian Herff, Lars Steinert. Predicting altcoin returns using social media, Dec 2018. https://www.researchgate.net/publication/329404921_Predicting_altcoin_returns_using_social_media
2. Ifigeneia Georgoula, Demitrios Pournarakis. Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices, 2015. https://pdfs.semanticscholar.org/05f3/7fcb95488402bb9e4d0d5a291e1338de41a6.pdf
3. Ifigeneia Georgoula, Demitrios Pournarakis. Using sentiment analysis to predict interday Bitcoin price movements, Nov 2017. https://www.researchgate.net/publication/321398204_Using_sentiment_analysis_to_predict_interday_Bitcoin_price_movements
4. Husnu Saner Narma, Jinwei Liu. Profile Analysis for Cryptocurrency in Social Media, Dec 2018. https://www.researchgate.net/publication/328737474_Profile_Analysis_for_Cryptocurrency_in_Social_Media
5. Connor Lamon, Eric Nielsen, E Fernández Redondo. Cryptocurrency Price Prediction Using News and Social Media Sentiment, 2017. https://www.semanticscholar.org/paper/Cryptocurrency-PricePrediction-Using-News-and-Lamon-Nielsen/c3b80de058596cee95beb20a2d087dbcf8be01ea
6. Connor Lamon, Eric Nielsen, E Fernández Redondo. Cryptocurrency Price Prediction Using News and Social Media Sentiment, 2017. http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf
7. Ignas Bagdonas. Predicting Bitcoin Price using Machine Learning, May 2018. https://www.researchgate.net/publication/326260718_Predicting_Bitcoin_Price_using_Machine_Learning
8. Ifigeneia Georgoula. Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices, May 2015. https://poseidon01.ssrn.com/delivery.php?ID=714022009114094068119069090111093126021037057034004075122020006076104020076117123117023061006041103116116085006002115070102123019027055089080085125075094090099086125063047077101068011097008075068090070125085024016105003120069097098031090088104095085083&EXT=pdf
9. Jermain Kaminski. Nowcasting the Bitcoin Market with Twitter Signal, Jab 2018. https://arxiv.org/abs/1406.7577
10. Jethin Abraham, Daniel Higdon. Predicting the Price of Bitcoin Using Machine Learning, 2018. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8374483&tag=1
11. John Nelson. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis, 2018. https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview
12. Evita Stenqvist, Jacob LÖnnÖ. Predicting Bitcoin price fluctuation with Twitter sentiment analysis, 2015. https://kth.diva-portal.org/smash/get/diva2:1110776/FULLTEXT01.pdf
13. Anique Akhtar. Machine learning for market trend prediction in Bitcoin, 2017. http://a.web.umkc.edu/aa95b/doc7.pdf
14. Young Bin Kim, Jurim Lee, Nuri Park. When Bitcoin encounters information in an online forum: Using text mining to analyses user opinions and predict value fluctuation, May 2017. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177630
15. Stuart Colianni, Stephanie Rosales. Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis, 2015. http://cs229.stanford.edu/proj2015/029_report.pdf
16. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.