A Corpus of BTC Tweets in the Era of COVID-19

Toni Pano, Rasha Kashef
Electrical, Computer, and Biomedical Engineering Department
Ryerson University
Toronto, ON, Canada
toni.pano@ryerson.ca, rkashef@ryerson.ca

Abstract: The coronavirus disease 2019 (COVID-19) outbreak has created a great challenge for many industries, including healthcare systems, E-commerce, and transportation. Cryptocurrencies, such as Bitcoin (BTC), is an alternative class of digital assets primarily used as a medium of exchange. In the financial market, data collected from social media such as tweets, posts, and can assist in an early indication of market sentiment. This has frequently been conducted on Twitter data. In particular, this paper provides a corpus of tweet text for Bitcoin-related tweets during the summer of the COVID-19 era. This dataset is publicly available covering a considerable period of roughly three months, to allow people with a similar interest in Twitter, Bitcoin, sentiment analysis, and the financial sector to perform research unimpeded.

Keywords—Bitcoin, Sentiment Analysis, Tweets, Covid-19

I. INTRODUCTION

Bitcoin is an electronic cash system that does not depend on financial institutions to keep track of its payments. Twitter Sentiment analysis is the automated process of analyzing tweets data and sorting it into positive, negative, or neutral sentiments. Using sentiment analysis to analyze opinions in Twitter data can generally help companies to get a better insight into customers' behaviors and interests.

As observed in [1-6], tweets' sentiments on BTC could serve as an indicator of Bitcoin prices. The collected corpus of the BTC tweets during the COVID-19 situation could predict the BTC trends and their correlation to its sentiment. Any observations of association may be of interest to the financial sector for price prediction or risk mitigation. The dataset is useful for forecasting the behavior of Bitcoin during pandemics, similar to COVID-19.

Access to specific tweets is usually restricted behind a paywall or by Twitter. Also, any publicly available tweets are limited to a historical depth of 7 days by Twitter's Search API [1] or are too general. Thus, this paper introduces the covid-19 BTC tweets dataset as a useful corpus for many research fields related to text mining, price prediction, financial analytics, and trend prediction.

Sentiments extracted from Twitter using Natural Language analysis may correspond with the behavior of Bitcoin prices. This is consistent with the observations of multiple works of research done by E. Stenqvist and J. Lönnö [1], O. Kraaijeveld and J. D. Smedt [2], T. R. Li et al. [3], S. Mohapatra, N. Ahmed, and P. Alencar [4], C. Kaplan, C. Aslan, and A. Bulbul [5], and A. Jain et al. [6]. The sentiments expressed in the dataset were intended to be correlated against the price of Bitcoin during COVID-19, specifically for observing any changes in correlation. Groups of research interest include bankers, investors, or other parts of the financial sector that deals with trend prediction or risk estimation. The dataset may be helpful to anyone interested in sentiment analysis or other natural language processing related to Bitcoin. The dataset is currently employed in a project for sentiment analysis and machine learning.

The rest of this paper is organized as follows: in section 2, the data description is provided. Section 3 summarizes the data specification. Experimental design and methods are discussed in section 4. Finally, conclusion and future directions are presented in section 5.

II. DATA DESCRIPTION

Tweets related to Bitcoin were collected by accessing Twitter's free REST APIs [7]. Specifically, the Search API was used to search for tweets created in the past [8,9]. A tweet scraper was built to collect the data and store it in a CSV file format. All relevant information for processing tweets on a timescale was included if it did not break Twitter's Developer Agreement [10]. In the interest of guarding user privacy, the username of each tweet's account was not recorded, and further preprocessing plans for the removal of usernames in text. Each tweet posted on a specific day and month has been stored in a file beginning with "month-day-2020". As the index of each row in a file increases, the timestamps of the recorded tweets decrease, starting around 23:59:59 and ending at 00:00:00.

III. DATA SPECIFICATION

All data in each file has been stored as a string, and certain data columns that don't apply are left blank. Each row of the dataset contains all related info for a single tweet, identified by the "id" column. The "time", "id" and "text" of any requested tweet comprises the first three columns, as a result of this labeled as the "requested tweet." The "time" column stores the timestamp of the tweet creation as "YYYY-MM-DD hh:mm:ss" in a 24-hour format. The "id" column stores the unique 19-digit number of each requested tweet.

978-1-7281-9615-2/20/$31.00 ©2020 Crown

The "text" column stores the text of the requested tweet. If the requested tweet was a retweet of someone else's tweet, then the next three columns are filled with the "original time", "original id", and "original text" of someone else's original tweet. In the case of the requested tweet being a retweet, the last two columns are filled by someone else's tweet info. If there requested tweet was not a retweet, then the last two columns are filled by the requested tweet. If the respective tweet text happens to be greater than 140 characters, the untruncated text fills the "full text" column with up to 280 characters [11]. If the respective tweet text happens to quote another tweet, a quotes object is formed and stored in a Javascript Object Notation (JSON) format, under the "quotes JSON" column. The quoted tweet is used to fill the "time", "id", and "full text" fields of the object, similar in function to the three columns of the CSV table with the same name. Should the text of a quoted tweet contain another quoted tweet, the quote JSON object will store another quote object under the "quote" field. This assumes that each tweet can only quote a single other tweet, as per Twitter's Retweet guide [12].

If this is not the case, then no "quote" field will be included. A quote object without that field denotes the end of a chain of such objects. Note that the quote objects seen in the figures have expanded the JSON text with whitespace for clarity. The JSON objects have been stored on a single line to save space in the dataset. Data specifications, as outlined by the Elsevier Journal's Data in Brief format, is shown in Table1. Sample Tweets are shown in Table 2. Fig.1. illustrates the structure of a quote object in the JSON form with no sub quotes. Fig.2 shows the structure of a quote object in the JSON form with two recursive sub quotes.

Table 1. Data Specification

Subject	**Computer Science**
Specific subject area	Research on Twitter Sentiment and Bitcoin Price Correlation
Type of data	CSV Table
How data were acquired	Requested from Twitter's Search API
Data format	Raw, strings
Parameters for data collection	Tweets were collected between 8:46:33 AM, May 22, 2020 and 23:59:59 PM, June 22, 2020. Requested tweets contained at least one of the following keywords: bitcoin, Bitcoin, bitcoins, Bitcoins, BTC, XBT, satoshi, #bitcoin, #bitcoins, #XBT, #BTC, $XBT, $BTC
Description of data collection	Timestamps, id, and the text of truncated, retweeted, full, and/or quoted text from each tweet
Data source location	Online, requires internet access
Data accessibility	https://drive.google.com/drive/folders/1zB__72h1O9N58diKzJ194B_Y-gaxycso?usp=sharing
Related research article	None

Table 2. Tweet rows for 5 sample tweets posted on June 22, 2020.

time	**id**	**text**	**original time**	**original id**	**original text**	**full text**	**quotes JSON**
2020-06-22 23:59:52	1275216...	Best ind...				Best ind...	{'time':20...
2020-06-22 23:59:51	1275216...	RT @IG...	2020-06-22 23:46:56	1275213...	#BlueLe...	#BlueLe...	
2020-06-22 23:59:50	1275216...	RT @Sm...	2020-06-22 6:55:33	1274959...	The fun...	The fun...	
2020-06-22 23:59:49	1275216...	@Alt...					
2020-06-22 23:59:47	1275216...	30 day...					

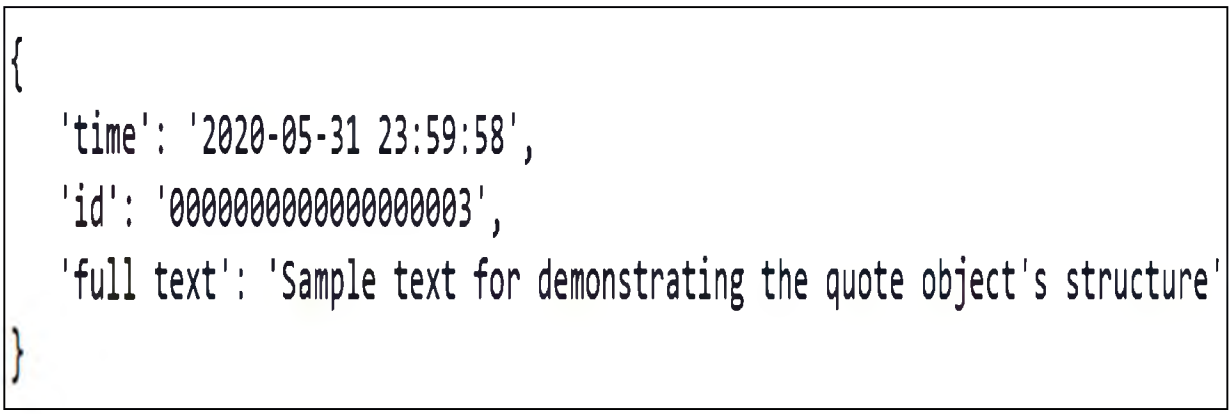

Figure 1. The structure of a quote object in JSON form with no sub quotes.

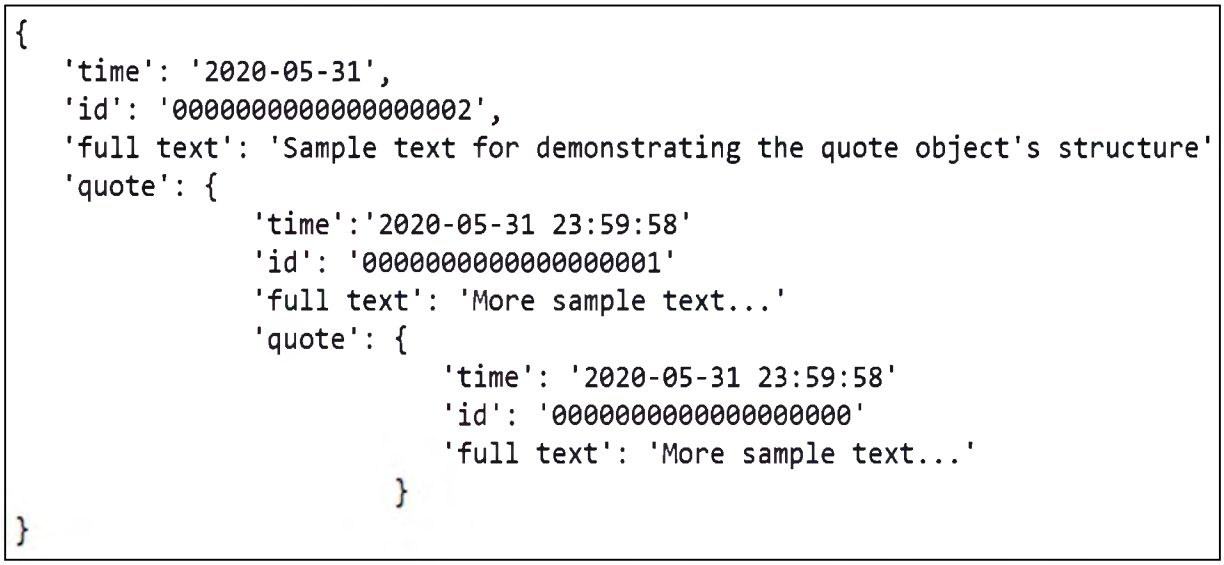

Figure 2. The structure of a quote object in a JSON format with 2 recursive sub quotes.

IV. Experimental Design and Methods

The Tweet collection was done by a script made in Python. This script used the Tweepy library [13] to requests tweets through Twitter's Search API [8,9]. The script was designed to collect all tweets it could from 00:00:00 to 23:59:59 of the previous day and store them in a single CSV file. Making our own script allowed us to make any unforeseen modifications needed for maintaining access to Twitter. The collection is intended to be done for each day from the end of May 2020 to the beginning of August 2020. All tweets from 8:46:33, May 22 to 23:59:59, June 22 have been collected so far. The dataset will continue to be updated until August 7, 23:59:59, once tweets have been gathered for all remaining days.

V. Business Adoption

The collected BTC tweets are useful datasets for many business applications to predict the BTC market movement directions. Financial analytics [14]-[16], anomaly detection [17]-[19], and machine and deep learning [20]-[22], and association mining [23],[24]..

VI. Conclusion and Future Directions

In this paper, we have provided the NLP community with a new twitter corpus collected in the era of COVID19. This corpus can be used in text mining, price prediction, financial analytics, and trend prediction. Future Directions include extending the corpus size to a more extended period to collect tweets in the entire COVID-19 period.

ACKNOWLEDGMENTS

This research has been supported by the Ryerson University, Faculty of Engineering and Architectural Science Undergraduate Research Opportunity Fund

REFERENCES

[1]. E. Stenqvist and J. Lönnö, 'Predicting Bitcoin price fluctuation with Twitter sentiment analysis', Dissertation,. Which journal it was published in ?, School of Computer Science and Communication, 2017.

[2]. O. Kraaijeveld and J. D. Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," Journal of International Financial Markets, Institutions and Money, p. 101188, Mar. 2020.

[3]. T. R. Li, A. S. Chamrajnagar, X. R. Fong, N. R. Rizik, and F. Fu, "Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model," Frontiers in Physics, vol. 7, Oct. 2019.

[4]. S. Mohapatra, N. Ahmed, and P. Alencar, "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments," Feb. 2020. Which journal it was published in ?

[5]. C. Kaplan, C. Aslan, and A. Bulbul, "Cryptocurrency Word-of-Mouth Analysis via Twitter," ResearchGate, 2018. [Online]. Available:

[6]. A. Jain, S. Tripathi, H. D. Dwivedi and P. Saxena, "Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, 2018, pp. 1-7, doi: 10.1109/IC3.2018.8530659. Conference Name:ACM Woodstock conference

[7]. https://developer.twitter.com/en/docs/basics/things-every-developer-should-know

[8]. https://developer.twitter.com/en/docs/tweets/search/overview

[9]. https://developer.twitter.com/en/docs/tweets/search/overview/standard

[10]. https://developer.twitter.com/en/developer-terms/agreement-and-policy

[11]. https://developer.twitter.com/en/docs/basics/counting-characters

[12]. https://help.twitter.com/en/using-twitter/how-to-retweet

[13]. https://github.com/tweepy/tweepy

[14]. Xue Tan and Rasha Kashef. 2019. Predicting the closing price of cryptocurrencies: a comparative study. In Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems (DATA '19). Association for Computing Machinery, New York, NY, USA, Article 37, 1–5. DOI:https://doi.org/10.1145/3368691.3368728.

[15]. Ibrahim, A.; Kashef, R.; Li, M.; Valencia, E.; Huang, E. Bitcoin Network Mechanics: Forecasting the BTC Closing Price Using Vector Auto-Regression Models Based on Endogenous and Exogenous Feature Variables. J. Risk Financial Manag. 2020, 13, 189.

[16]. Tobin T., Kashef R. (2020) Efficient Prediction of Gold Prices Using Hybrid Deep Learning. In: Campilho A., Karray F., Wang Z. (eds) Image Analysis and Recognition. ICIAR 2020. Lecture Notes in Computer Science, vol 12132. Springer, Cham. https://doi.org/10.1007/978-3-030-50516-5_11.

[17]. Kashef R., Gencarelli M., Ibrahim A. (2020) Classification of Outlier's Detection Methods Based on Quantitative or Semantic Learning. In: Fadlullah Z., Khan Pathan AS. (eds) Combating Security Challenges in the Age of Big Data. Advanced Sciences and Technologies for Security Applications. Springer, Cham. https://doi.org/10.1007/978-3-030-35642-2_3.

[18]. Kashef, Rasha, and Mohamed S. Kamel. "Towards better outliers detection for gene expression datasets." 2008 International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies. IEEE, 2008.

[19]. Rasha F. Kashef ; Proceedings of the KDD 2017: Workshop on Anomaly Detection in Finance, PMLR 71:43-55, 2018.

[20]. Kashef R. (2020) Adopting Big Data Analysis in the Agricultural Sector: Financial and Societal Impacts. In: Pattnaik P., Kumar R., Pal S. (eds) Internet of Things and Analytics for Agriculture, Volume 2. Studies in Big Data, vol 67. Springer, Singapore. https://doi.org/10.1007/978-981-15-0663-5_7

[21]. R. Kashef and M. S. Kamel, "Hard-fuzzy clustering: A cooperative approach," 2007 IEEE International Conference on Systems, Man and Cybernetics, Montreal, Que., 2007, pp. 425-430, doi: 10.1109/ICSMC.2007.4413889.

[22]. A. Ibrahim, D. Rayside and R. Kashef, "Cooperative based software clustering on dependency graphs," 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, 2014, pp. 1-6, doi: 10.1109/CCECE.2014.6900911.

[23]. El-Sonbaty, Yasser, and Rasha Kashef. "New Fast Algorithm for Incremental Mining of Association Rules." ICEIS (1). 2004.

[24]. Sonbaty, Yassar El, and Rasha F. Kashef. "NBP: Negative Border with Partitioning Algorithm for Incremental Mining of Association Rules." (2004).