Date of publication: February 14, 2022

Using Real-time Data to Analyze General Sentiment about Crypto-currency

PETAR GRIGOROV[1], MARIAH MARIN[2], MARC MEIER-DORNBERG [2], KHILNA RAWAL[2], GABRIELLE STONEY[3]

[1]Department of Physical Sciences, College of Arts and Sciences, Embry-Riddle Aeronautical University, Daytona Beach, FL
[2]Department of Mathematics, Colleg of Arts and Sciences, Embry-Riddle Aeronautical University, Daytona Beach, FL
[3]Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL

ABSTRACT Crypto-currency is a digital currency designed to work as means of exchange through a computer network. Unlike the stock market, crypto is decentralized, meaning that it does not rely on a central authority (such as a government or bank) to uphold or maintain it. In contrast with actual stocks, crypto-currency is a highly volatile and sentiment-driven form of investment. While stock market indices depend on political and socioeconomic factors such as interest rates, current events, government upheaval, exchange rate fluctuations, natural calamities and so on, the demand for a particular crypto depends mostly on the overall attitude of investors towards it. Social media platforms such as Twitter are among the main sources where information about these sentiments can be extracted. The scope of the project is to use the free Twitter API to collect tweets over a given period of time and use this data to analyze the general sentiment about crypto-currency. In addition to that, the price movements of a crypto-currency (e.g. bitcoin) will be collected over that same period of time. The goal is to predict the price movement for the hours that follow using the sentiment scores mentioned previously. Data will be integrated from two different real-time sources, namely the Twitter API and a crypto-currency data provider. The features will then be extracted and prepared for analysis. An additional potential scope for this project is using cloud capabilities to reliably collect real-time data over a longer time span.

INDEX TERMS Crypto-currency, Stocks, Investment, Social media, Database, Real-time data, Time series prediction

I. INTRODUCTION

THE Twitter API can be used to retrieve and analyze Twitter data using computer programming, as well as build for the conversation on Twitter. Over the years, the Twitter API has grown by adding additional levels of access for developers and academic researchers to be able to scale their access to enhance and research the public conversation. The recently released Twitter API v2 includes a modern foundation, new and advanced features, and quick on-boarding to *Essential* access that allows to retrieve up to 500k tweets per month or 25 requests per 15 minutes. There are also several other APIs for stocks and crypto currency such as CoinAPI, which offers a free plan with up to 100 requests per day.

A. RELEVANT LITERATURE

There have been numerous studies that used Machine Learning models for time series forecasting with the purpose of predicting stock prices or values of crypto currencies, and to conduct sentiment analysis either using Twitter APIs or other media sources. One paper by Ikhlaas Gurrib and Firuz Kamalov, both faculty of the Canadian University of Dubai, proposed a new method to predict values of cryptocurrencies such as bitcoin using sentiment analysis and linear discriminant analysis (LDA). [1] The model essentially trains an LDA classifier that uses information from real-time bitcoin prices and sentimental news/media headlines to forecast bitcoin prices for the following day.

A thesis study done by Roderick Karlemstrand and Ebba Leckstrom from the KTH Royal Institute of Technology attempted to use a Machine Learning time-series model to predict stock prices, this time using Twitter's API to collect data. The model is based on a neural network that is trained with historical stock values and attributes that have been extracted from posts on Twitter. These attributes represent sentiment scores, retweets, followers, etc. [2] The analysis was conducted using a rule-based sentiment analysis tool called *Valence Aware Dictionary and sEntiment Reasoner* (VADER).

A third paper by Chamrangar et al. also used Twitter sentiments to predict price fluctuations of *ZClassic*, an alternative form or cryptocurrency. Each tweet is extracted and classified as either positive, neutral, or negative, and later compiled into two time-series sentiment indices (one weighted and one unweighted). The two indices were trained on an Extreme Gradient Boosting Regression Tree Model. The significance of this paper is that it is the first academic proof of the concept of how powerful social media is in influencing price movements of the highly volatile cryptocurrency market. [3] The final source reviewed for this study is a paper by Loginova et al., which uses similar sentiment-based analysis as the other papers, except the dataset consists of data from Reddit, Bitcointalk and CryptoCompare instead of social media.

II. SYSTEM DESIGN

CLOUD platforms provide numerous advantages over a traditional on-premise architecture. Services can be activated easily and often without the need to install them. Cloud native-services are maintained by cloud service providers in a robust manner so that efforts for technical operation and maintenance can be reduced. The most important advantage of cloud computing for this project is the possibility to collect data reliably over the course of several weeks.

The project team has decided to use the Google Cloud Platform (CGP) for retrieving and storing data, training the model, and computing the prediction. The system architecture that takes advantage of GCP's cloud-native services will be presented in the following.

FIGURE 1. System Overview

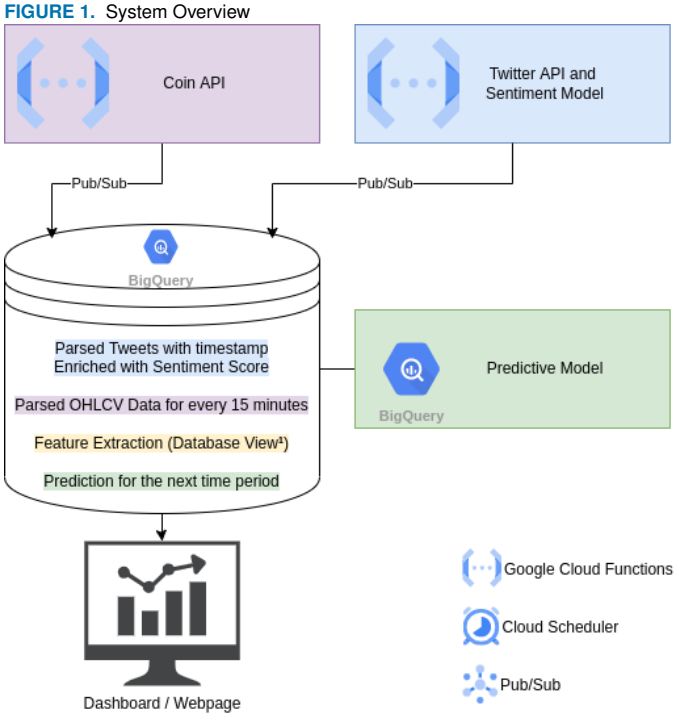
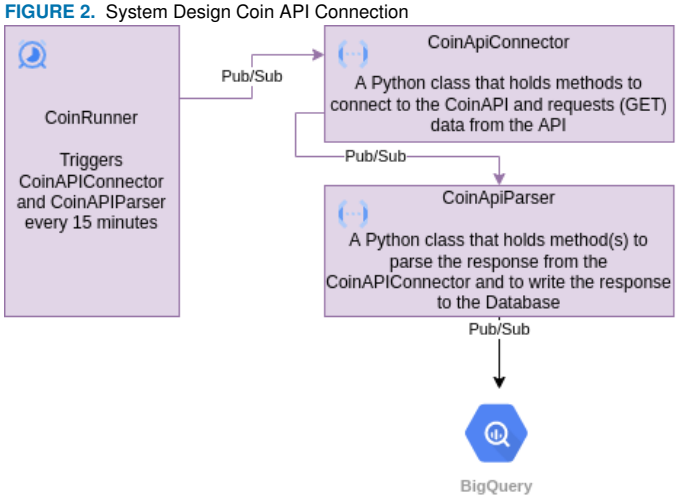


Figure 1 shows a high-level system overview. Data on cryptocurrency price movements is collected from an API provided by Coin API and tweets from Twitter are collected using the Twitter API. A sentiment score is computed for every incoming tweet. Both data feeds are stored in BigQuery, the data warehouse solution in GCP [5]. A predictive model is trained on the collected data to forecast the price movements for the respective succeeding time period. A prediction is updated periodically so that a system user can make investment decisions in real-time. The results are ultimately displayed on a dashboard or web page.

A. API CONNECTIONS

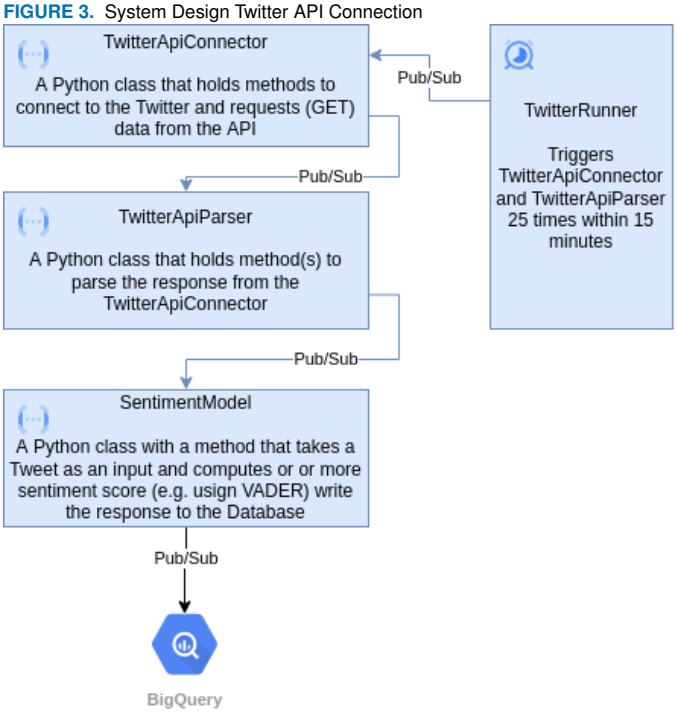
FIGURE 2. System Design Coin API Connection



Both API connections, the Coin API (Figure 2) as well as the connection to the Twitter API (Figure 3), follow the same architectural pattern. Both connections use Google Cloud Functions, a lightweight solution in GCP, to deploy serverless functions [6]. The Cloud Functions contain Python code, that produces the desired outcome. There are two Cloud Functions for the connection to the Coin API. The first Cloud Function (CoinApiConnector) connects to the API and outputs the API's response string. The second Cloud Function (CoinApiParser) takes the input from the CoinApiConnector and parses it into the desired format. The Twitter API connection follows the same approach. In addition to that, there is a third Cloud Function (SentimentModel) that takes the parsed tweets and computes a sentiment score. The output of each input stream's last Cloud Function (CoinApiParser and SentimentModel) is written to BigQuery.

Pub/Sub is Google Cloud's asynchronous messaging service [7]. Pub/Sub acts as a buffer and messaging queue between the Cloud Functions so that an overload of the Cloud Function's processing power can be avoided. Each Cloud Function publishes its output to a dedicated Pub/Sub queue from where it is consumed by the next Cloud Function for further processing.

FIGURE 3. System Design Twitter API Connection



The CoinApiParser and the SentimentModel are the last Cloud Functions in their respective data stream before the output data can be written to BigQuery. Both of these components publish their result in a JSON format to Pub/Sub where they can then be written to BigQuery [8].

Google Cloud's implementation of a cron service, Cloud Scheduler [9], is used to schedule the API calls. Cloud Scheduler triggers the Cloud Functions periodically via messages on a dedicated Pub/Sub queue. The API calls are triggered, in a way that makes use of the upper limit of allowed API calls per time period (e.g., one API call every 15 minutes to CoinAPI).

B. DATA STORAGE AND DATA PREPARATION

As mentioned previously, data collected by the Cloud Functions is stored in BigQuery. BigQuery has a landing zone for the data ingested by the Coin API stream and for the data ingested by the Twitter API stream. One table in the landing zone contains one record with cryptocurrency price data for the respective period. Another table contains one tweet and an attached sentiment score per record. These two tables need to be combined and consolidated before the data can be used for model training.

There are two possible approaches to prepare the input tables for model training. Tables can be written periodically (e.g., after every time the CoinAPI stream added a record to the landing zone) or database views can be used. The second option is preferred since development efforts can be kept low by using database views[1].

The following list gives an overview of possibly relevant features where 'p' is a time period and 'p-1' is the previous time period. The Twitter API and the Coin API provide many other potential features that can be added to the list.

- Number of tweets during time period p
- Average Sentiment during p
- Average Sentiment during (p-1)
- Average Sentiment during (p-2)
- ...
- Average Sentiment during (p-n)
- price close after p
- price close after (p-1)
- price close after (p-2)
- ...
- price close after (p-n)
- price p - price p-1
- price p-1 - price p-2
- ...
- price p-n-1 - price p-n-2

C. MODEL

BigQuery ML allows developers to create Machine Learning models with SQL syntax [10]. Developers can create, train, and store the model in BigQuery itself. This is the preferred way of creating a model for this project as it allows the team to focus on the model itself rather than additional interfaces to BigQuery. The model is used within BigQuery to generate predictions that can be read from a consuming application such as a dashboard.

REFERENCES

[1] Gurrib, I., amp; Kamalov, F. (2021, December 15). Predicting bitcoin price movements using sentiment analysis: A machine learning approach. Studies in Economics and Finance. Retrieved February 1, 2022, from https://www.emerald.com/insight/content/doi/10.1108/SEF-07-2021-0293/full/html?skipTracking=true

[2] Karlemstrand, R., amp; Leckstrom, E. (2021, May 4). Using Twitter attribute information to predict Stock Prices. Using Twitter Attribute Information to Predict Stock Prices. Retrieved February 1, 2022, from https://arxiv.org/pdf/2105.01402v1.pdf

[3] Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., amp; Fu, F. (2019, July 10). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. Frontiers. Retrieved February 1, 2022, from https://www.frontiersin.org/articles/10.3389/fphy.2019.00098/full

[4] Loginova, E., Tsang, W. K., van Heijningen, G., Kerkhove, L.-P., amp; Benoit, D. F. (2021, November 18). Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data - machine learning. SpringerLink. Retrieved February 1, 2022, from https://link.springer.com/article/10.1007/s10994-021-06095-3

[5] Google Cloud. BigQuery. Retrieved February 9, 2022, from https://cloud.google.com/bigquery

[6] Google Cloud documentation. Google Cloud Functions. Retrieved February 9, 2022, from https://cloud.google.com/functions/docsdocs

[7] Google Cloud documentation. What is Pub/Sub? Retrieved February 9, 2022, from https://cloud.google.com/pubsub/docs/overview

[1]However, this approach could cause higher costs or even performance problems due to the increased query complexity. Costs and performance have to be closely monitored during development. The development team may decide to discard the option of using database views.

[8] Google Cloud documentation. Google-provided streaming templates. Retrieved February 9, 2022, from https://cloud.google.com/dataflow/docs/guides/templates/provided-streaming
[9] Google Cloud documentation. Cloud Scheduler. Retrieved February 9, 2022, from https://cloud.google.com/scheduler
[10] Google Cloud documentation. What is BigQuery ML? Retrieved February 9, 2022, from https://cloud.google.com/bigquery-ml/docs/introduction
[11] Diagrams created with draw.io