CS475 Final Report: Team J.A.R.V.I.S. (12)
Cryptocurrency Price Movement Prediction from Tweets

Eugene Lee
KAIST
School of Computing
eugene@kaist.ac.kr

Webi Dabuse
KAIST
School of Computing
webby@kaist.ac.kr

Jaeryoung Ka
KAIST
School of Computing
jaeryoung.ka@kaist.ac.kr

Abstract

Cryptocurrencies recently became new, wildly-acknowledged monetary assets, regarding which lots of interactions similar to traditional stock is occurring. Given that much stock price prediction research revealed unique insights to understanding how the stock market works, it's worth investigating the cryptocurrency context with similar approaches. However, little on cryptocurrencies are investigated yet despite its surging attention and bigger acknowledgements as monetary assets. In this work, we leveraged existing research about predicted stock price fluctuation from social network corpora to apply multiple generative models from (Xu and Cohen, 2018) towards Bitcoin and Ethereum contexts to make temporally-dependent movement predictions. Our replications succeeded to achieve relatively good accuracies for both cryptocurrencies. We also revealed that price movement patterns of cryptocurrencies quite differ from that of traditional stocks.

1 Introduction

Cryptocurrencies have become a new black in social networks. Our pipeline shows that the number of tweets mentioning either `Bitcoin` or `Ethereum` surged up by 5.24 times, comparing January 2017 and January 2021. Given the price of `Bitcoin` has been multiplied by 4.03 during the same period, the hot attention from users would not be a coincidence.

There exist much research to leverage various corpora online to predict stock price fluctuations. (Xu and Cohen, 2018) and (Pagolu et al., 2016) leverage social media text, whereas (Hu et al., 2018) utilizes news corpora. Such predictions often enable governments to monitor macroeconomy, researchers to better understand the hidden correlations between price and other factors (e.g., sentiments of investors), and investors to reap more profit.

However, despite the benefits and hot attentions, answers to how to apply such methods to cryptocurrency prices have been underinvestigated. In this research, we replicated a few generative models from (Xu and Cohen, 2018) to make temporally-dependent predictions in the stochastic cryptocurrency market.

Given three shared characteristics: (1) stochasticity of markets, (2) temporal-dependency of predictions, and (3) very chaotic data, we expected that the model for stock prices would function well for cryptocurrency as well. Our findings suggest that the models behave quite decently in the cryptocurrency context, though there are quite different patterns where lots of future investigations can be made.

2 Approach

To predict the movement of a target cryptocurrency price $\boldsymbol{c} \in \{ETH, BTC\}$ on a trading day $\boldsymbol{d}$, we use the market information comprising relevant twitter corpora and historical prices. For a prediction on a trading day $\boldsymbol{d}$, we use the market information that was collected in the lag $[\boldsymbol{d} - \boldsymbol{\Delta d}; \boldsymbol{d} - \mathbf{1}]$ where $\boldsymbol{\Delta d}$ is a fixed lag size. This approach will estimate the binary movement of stocks where 1 denotes rise and 0 denotes fall.
Binary prediction $\boldsymbol{y} = \mathbf{1}(\boldsymbol{p}_{\boldsymbol{d}}^{\boldsymbol{c}} > \boldsymbol{p}_{\boldsymbol{d}-\mathbf{1}}^{\boldsymbol{c}})$, where $\boldsymbol{p}_{\boldsymbol{d}}^{\boldsymbol{c}}$ denotes the adjusted closing price adjusted for corporate actions affecting stock prices.

The lag size determines the number of days used to collect market information. This is used to incorporate the prediction of other days in the lag of trading day $\boldsymbol{d}$ to simulate the prediction targets close to $\boldsymbol{d}$. Our approach uses the advantage of multi-task learning (Caruana, 1998) in the collection of market information in the lag $[\boldsymbol{d} - \boldsymbol{\Delta d}; \boldsymbol{d} - \mathbf{1}]$.

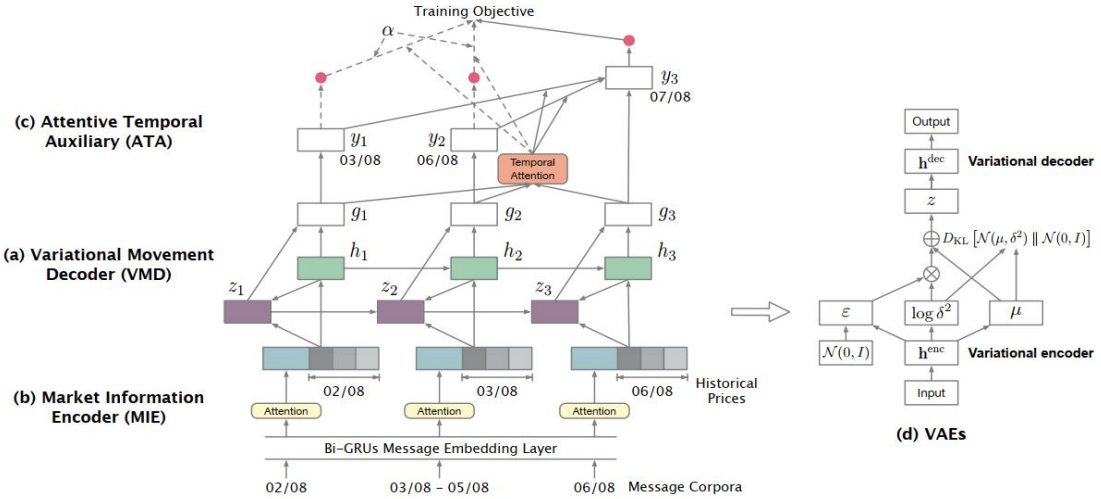

Figure 1: The architecture of StockNet using main target of 07/08/2012 and the lag size of 5 for illustration.

3 Data and Experiments

3.1 Data

The original paper (Xu and Cohen, 2018) selected two-year price movements from 01/01/2014 to 01/01/2016 of 88 stocks to target. In our replication, we selected three-year price movements from 01/01/2017 to 31/12/2019 of cryptocurrency coins. We used Bitcoin and Ethereum coins as our target stocks. The market information representing both stocks is collected from Twitter and yahoo finance. We scraped about 30 million tweets on Bitcoin and Ethereum by using Twint – an advanced Twitter scraping tool written in Python that allows scraping Tweets from Twitter profiles without using Twitter's API. Since we planned to collect a lot of data within a short amount of time, we could not use the official Twitter API. We preprocess the tweet texts to get better quality data. We tokenize the text and add different methods of treating information from emojis, hyperlinks, hashtags, the "$" sign, and the "@" identifier.

The other market information we used is historical prices, which we collected from yahoo finance by using their official API. This price data is normalized based on the previous day's closing price. That means, today's stock price is recorded by the increments or decrements it has from yesterday's closing price. To label the price data, we used the same approach on the original paper. Since we aim at the binary classification of coin price changes identifiable from social media, we set two particular thresholds, 0.5% and 0.55% Samples with the movement percents $\leq -0.5\%$ and 0.55% are labeled with **0** and **1**, respectively.

3.2 Model

Our model is the same as the original paper's model. As displayed in figure 1, the model comprises three main components in a bottom-up fashion:

1. Market Information Encoder (**MIE**)
2. Variational Movement Decoder (**VMD**)
3. Attentive Temporal Auxiliary (**ATA**)

Market Information Encoder

MIE encodes information from social media and cryptocurrency prices to enhance market information quality, and outputs the market information input $\mathbf{X}$ for the next component – Variational Movement Decoder. Each temporal input is defined as $\boldsymbol{x_t} = [\boldsymbol{c_t}; \boldsymbol{p_t}]$ where $\boldsymbol{c_t}$ and $\boldsymbol{p_t}$ are the corpus embedding and the historical price vector, respectively

Variational Movement Decoder

The purpose of VMD is to recurrently infer and decode the latent driven factor $\mathbf{Z}$ and the movement $\mathbf{y}$ from the encoded market information $\mathbf{X}$.

Attentive Temporal Auxiliary

ATA component integrates temporal loss through an attention mechanism for model training. With the acquisition of a sequence of auxiliary predictions in the given lag, two-folded auxiliary effects are incorporated into the main prediction and the training. In this component, $\boldsymbol{\alpha}$ represents the overall auxiliary effects on the model training. We will

discuss how this hyperparameter affects the learning process of our model in the discussion section.

3.3 Experiments

Regarding the data, as discussed earlier, we use cryptocurrency market information instead of 88 stocks in the original paper. After preprocessing the tweets and adjusting the price data, we train the model in the following conditions. We use a 5-day lag window for sample construction of a target day prediction and 32 shuffled samples in a batch. The word embedding size was set to 50 and we used GloVe. We train the model with an Adam optimizer with an initial learning rate of 0.001. We experimented with different α, which is the overall auxiliary effects on the model training, for different variants of the proposed model, StockNet.

These variants are:

1. **Technical Analyst**: a variant that uses only historical prices.
2. **Fundamental Analyst**: a variant that uses only tweet information.
3. **Independent Analyst**: Independent of temporal auxiliary targets.
4. **HedgeFund Analyst**: Combination of both Technical Analyst and Fundamental Analyst. It uses both historical prices and tweet information.
5. **Discriminative Analyst**: The discriminative StockNet directly optimizing the likelihood objective.

4 Results

Because stock market predictions are challenging and small improvements can lead to large gains, the accuracy of 56% is reported as a sufficient result (Nguyen and Shirai, 2015). The original StockNet paper reports HAN (Hu et al., 2018) with the best accuracy of 57.64 and TSLDA (Nguyen and Shirai, 2015) with the best MCC of 0.065382. HAN is a hierarchical attention neural net that only uses tweets whereas TSLDA is a generative model that uses news and prices. Although the five baseline models aren't fully comparable, as we didn't re-run them for our crypto dataset, they are still referenced here.

We apply the model presented in the paper to cryptocurrencies, namely Bitcoin and Ethereum. Using our crypto dataset, the DISCRIMINATIVEANALYST achieves the best performance in accuracy of 59.55 and in MCC of 0.173984 at an α of 0.1. This is a 2% increase in accuracy and 116% increase in MCC versus the paper's HEDGEFUNDANALYST variant at an α of 0.5. Additionally, this is a 3% increase in accuracy to TSLDA and 166% increase in MCC to HAN.

Similar to the paper, we explore how the temporal auxiliary effects alter the model's performance. This objective level temporal auxiliary variable can be regarded as a denoising regularizer. The paper mentions that α helps to filter market sources in the lag as per their respective aligned auxiliary movements (Xu and Cohen, 2018). For the HEDGEFUNDANALYST variant, it allows the model to better generate stock movements from latent driven factors, consistent with recent research in Variational Auto-Encoders.

When using the cryptocurrency dataset, we see that the model does not linearly benefit in accuracy and MCC from the changes in α (Figure 3). This trend is similar to the paper's, but differ in the best-performing variant type. In our cryptocurrency dataset, the best performance came from the DISCRIMINATIVEANALYST rather than the HEDGEFUNDANALYST. The DISCRIMINATIVEANALYST with α of 0.1 has the largest accuracy of 59% and the HEDGEFUNDANALYST with α of 0.5 has the largest accuracy of 54%.

The HEDGEFUNDANALYST is the combination of the TECHNICALANALYST and FUNDAMENTALALANALYST, 57.89 and 56.32 in accuracies respectively (Figure 2). From these results, the HEDGEFUNDANALYST should be able to leverage both variants to improve performance; however, underperforms both with the best accuracy at 54.21 (Figure 2). We believe this is because cryptocurrency movements are inherently different from those of stocks. Stocks or cryptocurrencies in the datasets are treated as a collection of data and act as one movement. Cryptocurrencies are much more volatile than stocks, thus small movements in stocks may be much easier for the generative model to predict. For crypto, this may be the reason that the discriminative model outperforms the generative model. The stocks also have a larger dataset (88 stocks vs 2 crypto) and the training range for crypto include a very volatile era in price movements, much different from the development and test date ranges.

StockNet Variations	Paper's Result		Our Result	
	Acc.	MCC	Acc.	MCC
TECHNICALANALYST	54.96	0.016456	57.89	0.141954
FUNDAMENTALANALYST	58.23	0.071704	56.32	None
INDEPENDENTANALYST	57.54	0.036610	46.84	-0.02866
DISCRIMINATIVEANALYST	56.15	0.056493	**59.55**	**0.173984**
HEDGEFUNDANALYST	**58.23**	**0.080796**	54.21	0.025192

Figure 2: Performance of StockNet variations in accuracy and MCC on stock dataset and crypto dataset

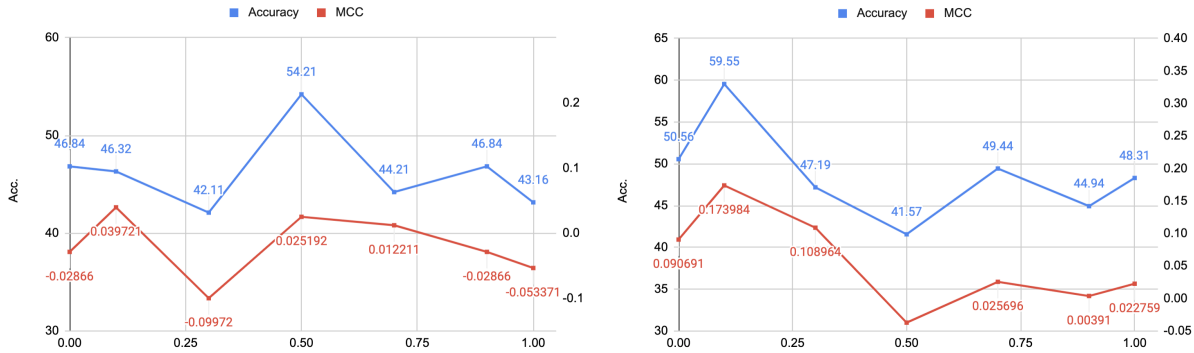

Figure 3: (Left) Performance of HEDGEFUNDANALYST with varied α. (Right) Performance of DISCRIMINATIVEANALYST with varied α.

5 Discussions

Our largest modification to the paper is using cryptocurrency rather than stocks in the model. We further improve the fetching and pre-processing of tweets to better handle emojis and misaligned strings. From our observations, the Twitter data for cryptocurrency is very chaotic; therefore, randomly dropping 60% of the tweets per day resulted in better accuracy and more efficient computations.

Limitations still exist in our preprocessing pipeline. As the tweets are very messy, scammers and bots aren't removed properly. Furthermore, the dates of prices were pulled in UTC and the tweets in KST so there exists challenges in perfectly aligning these two data sets. Lastly, the effects of the temporal auxiliary are more complex than we assumed. It would be worthwhile to explore all of these limitations in future works.

References

Rich Caruana. 1998. *Multitask Learning*, pages 95–133. Springer US, Boston, MA.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction.

Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices.