Predicting Cryptocurrency Price Movements Based on Social Media

Van Minh Hao[1], Nguyen Huynh Huy[1], Bo Dao[2], Thanh-Tan Mai[3], Khuong Nguyen-An[1, (✉)]

[1]Faculty of Computer Science and Engineering, University of Technology, VNU-HCM, Vietnam
{1510901, 1511252, nakhuong}@hcmut.edu.vn

[2]School of Information Technology, Deakin University, Australia.
bo.dao@deakin.edu.au

[3]Department of Mathematics and Statistics, Quy Nhon University, Vietnam.
maithanhtan@qnu.edu.vn

***Abstract*—Predicting cryptocurrency price movements is a challenging task due to the highly stochastic nature of the market. This paper exploits features from social media, combining with the historical price to build an accurate model for predicting trending of the Bitcoin, the most popular cryptocurrency these days. The novelty of this work is introducing a new feature called *interaction* which helps improve the model's performance significantly. Our approach shows a very promising result, which outperforms other recent works by a large margin.**

***Index Terms*—Bitcoin, cryptocurrency, market prediction, sentiment analysis, machine learning, data analysis**

I. Introduction

Recent years have witnessed an explosion in the cryptocurrency market. More and more investors join this market due to its high profit. On July 2, 2019, six largest cryptocurrencies achieved a total value of US $264.6 billion in terms of market capitalization[1], in which Bitcoin only contributes US $199.9 billion. At the same time, Bitcoin price experiences high volatility rate making many investors confused. In one week, from June 19, 2019 to June 27, 2019, value of a unit of Bitcoin increased US $4,700, achieved US $13,700. Within the same duration between June 27, 2019 and July 2, 2019, the Bitcoin price went down from US $13,757 to US $9,790. These phenomena raise an important question whether we could help investors reduce the uncertainty in their investment decisions. This question is even more challenging when considering these fluctuations not only depend on historical prices but also other factors, especially the overall mood of society about cryptocurrency. Nowadays, the emergence of social media makes this mood data largely available. In this paper, we take some steps by leveraging these information from online social network platforms to build a model for predicting daily Bitcoin price movements based on Gradient Boosting algorithm. Our work only considers Bitcoin and its daily change at Coinbase[2], one of the most reputable cryptocurrency exchanges. Although the standard procedure for this kind of prediction has been widely applied, this study adds a new ingredient:

We propose a novel feature, named ***interaction*** for Bitcoin market prediction.

Besides that, we release our dataset[3] collected for this study. We believe that this new dataset would be useful for the community for further studies on this topic.

The rest of our paper is organized as follows. Section II reviews some recent works on cryptocurrency price movements prediction, especially those using information from online sources as features. Section III recalls the background knowledge of Gradient Boosting algorithm and VADER, a method to analysis sentiment in text documents. These are building blocks for out method. Section IV describes our prediction procedure, features for model, especially our new ***interaction*** feature. Section V reports all experimental results, especially shows the effectiveness of our novel feature in contributing to model performance. Finally, Section VI summaries the paper and envisions some future research directions.

II. Related works

Over the last few years, many researchers tried to develop various ways for Bitcoin market prediction. Some studies used techniques from time-series forecasting [9], [10], [11]. Others viewed the problem as a supervised learning problem [12], [13], [14]. A common theme of these studies is to use two main sources of information: historical prices and Bitcoin-related information from the Internet. Due to the scope of our paper, this section will mainly review some recent notable works which leverage online social signals in predicting procedure.

Abraham *et al.* [1] proposed a method for predicting changes in cryptocurrency price based on Twitter data and

[1] https://bitinfocharts.com/

[2] https://www.coinbase.com/

[3] https://github.com/minhhao97vn/BTCDataset

978-1-7281-4723-9/19/$31.00 ©2019 IEEE
DOI 10.1109/ACOMP.2019.00016

Google Trends data. The result of analyzing input data showed that both tweet volume and Google Trends were highly correlated with price. A multiple linear regression models was used to predict cryptocurrencies price with these input features.

In the study of Mai *et al.* [2], the authors showed that the movements of Bitcoin price related to social media information. Comments and replies on Twitter and Bitcointalk forum were collected for analyzing task. The result showed that social media information, especially Bitcointalk, had a high impact on Bitcoin price.

In [3], Kim *et al.* presented a method for predicting fluctuations in cryptocurrencies price and number transactions. The authors collected user comments, replies on forums and used Granger causality test to analyze the correlation between social media information and price. Model AODE (Average One-Dependent Estimators) was used to predict fluctuations.

McNally *et al.* [4] predicted Bitcoin price and price movements using historical price. Recurrent Neural Network model (RNN) and Long Short Term Memory (LSTM) model with Bayesian optimization were implemented for predicting task. An ARIMA (Autoregressive Integrated Moving Average) time series model is also developed for performance comparison purposes with the neural network models. The LSTM outperformed the RNN and deep learning models had a better prediction than ARIMA.

In [5], I. Georgoula *et al.* proposed to use time series analysis to learn the relationship between Bitcoin price and economics, technological variables and sentiment from Twitter. In sentiment analysis task, Support Vector Machines (SVM) was used to measure the Twitter sentiment ratio on a daily basis. In time series analysis task, the authors estimated two regression model: Ordinary Least-Square (OLS) Estimates for short-run and Vector Error-Correction Model (VECM) for long-run. The results showed that Twitter sentiment ratio had a positive short-run impact on Bitcoin price. Moreover, the price of Bitcoin was found to be positively affected by the number of Wikipedia search queries.

Using a similar approach to [1], Matta *et al.* [6] presented a method for analyzing cross-correlation between Bitcoin price and tweet sentiment, views for Bitcoin on Google Trends. The analysis showed that Google Trends SVI had a high correlation with price.

III. Background

A. Cryptocurrency trading

By May 2019, there are approximately 250 cryptocurrency exchanges around the world. There are two types of cryptocurrency exchange: centralized exchange (CEX) and decentralized exchange (DEX). The characteristics of each type are as follows:

- ***Centralized EXchange*** (CEX): The fundamental of a CEX is that it is the intermediary between two investors. Each exchange controls the entire system and activities for all transactions done on its platform. Every investors joining an exchange will have an account of cryptocurrency exchange (or often called *wallet*), which contains all money of investors. Two investors will interact through the Exchange to buy/sell cryptocurrencies. Some of the most notable Centralized Exchanges are Coinbase, Binance, Kraken, etc.
- ***Decentralized EXchange*** (DEX): On comparison to a CEX, a DEX does not use any intermediary system to work. The system will be fully automated and based on a peer-to-peer model. Like the CEX above, the DEX also uses *wallet* to save all investor's asset. Transactions of converting between cryptocurrency and fiat money are not allowed for this type of exchange. Some of the most popular Decentralized Exchange are IDEX, Waves, Bitshares, etc.

When joining this market, investors should know some basic rules as follows:

- Each kind of exchange has different kind of transactions. For instance, while the CEXs allow investors to convert between cryptocurrency and fiat money, the DEXs do not allow this.
- A unit of cryptocurrencies on different exchanges does not have the same value. Each exchange sets trading pairs rate between each pair of cryptocurrencies. This rate varies based on many factors.
- Each exchange has different transaction fee. Investors should consider this when joining an exchange.
- Transactions are allowed to execute at any time of the day. Unlike financial stock market, there is nothing called trading session on cryptocurrency market since it opens 24/7.

Besides those rules, investors need to gain knowledge about trading strategies to earn profit and reduce risk in their investment decisions. Here we give some tactics through simple scenario when trading cryptocurrencies:

- Trading on multiple exchanges: There are cryptocurrencies pairs BTC/ETH that have trading rate 1/2.2 on exchange A and trading rate 1/2.0 on exchange B. An investor has 1 BTC on exchange A. This guy could trade by using strategy as follows. At the beginning, he/she trades 1 BTC to receive 2.2 ETH on exchange A. Then he/she transfers this 2.2 ETH from exchange A to exchange B with the fee 0.05 ETH. He/she continues to trade the rest 2.15 ETH on exchange B to obtain 1.075 BTC. The final profit above investor gains from this investment is 0.075 BTC.
- Cycle arbitrage: This tactic leverages immediate trading rates on an exchanges to make profit. For instance, trading rates on exchange A of three cryptocurrencies pairs BTC/ETH, ETH/LTC and LTC/BTC at the moment are 1/1.1, 1/1.2 and 1/0.8, respectively. If one has 1 BTC at the beginning, after trading a cycle of these three cryptocurrencies, he or she can earn a profit of 0.056 BTC. The most important concern here is that the transfer takes time while the trading rates changes fast based on the number of transactions. During this time, the arbitrage

opportunity may already be gone. Therefore, one need to develop automated tools to detect these behaviors at high speed to make as much profit as possible.

B. Gradient Boosting algorithm

Gradient Boosting (GB), invented by Friedman [20] is one of the most powerful predictive algorithms these days. This subsection summaries GB mainly based on [21]. Assume that we have a dataset $(x_i, y_{)}{}_{i=1}^{N}$ with N observations, each x_i has d dimensions. The problem is to find an estimate $\hat{f}(x)$ that approximates "unknown true function" f representing the dependence between x and y: $x \overset{f}{\mapsto} y$, under constraint chosen loss function $\mathcal{L}(y, f)$ is minimized. Many traditional algorithm solves this task by restricting search space from function search space to a parametric family of function $f(x, \theta)$. This would then convert problem of function approximation to problem of parameter estimation.

$$\hat{f}(x) = f(x, \hat{\theta}),$$

where

$$\hat{\theta} = \arg\min_{\theta} E_{x,y}(\mathcal{L}(y, f(x, \theta))).$$

At this stage, we can solve above problem by using standard numerical optimization algorithms, such as Gradient Descent, Stochastic Gradient Descent, etc.

The fundamental difference between Gradient Boosting and many traditional machine learning algorithms which have formulation like above is that the optimization step is done on function search space instead parameter search space. Friedman parameterized the estimated function $\hat{f}$ as sum of many function increments (or called with another name: "boosts").

$$\hat{f}(x) = \sum_{j=0}^{M} \hat{f}_j(x),$$

where:

- m is the number of iteration,
- $\hat{f}_0$ is the initial function,
- $\hat{f}_j$ is the function increment.

Step-size ρ will be optimally chosen for each iteration *t*-th.

To specify notation for Gradient Boosting algorithm, a parameterized function is introduced as a base-learner $h(x, \theta)$. One can choose in advance different base-learner such as linear models, decision tree model, etc. Finally, we need to solve below optimization model.

For function estimate at $t-$th iteration:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t),$$

where

$$(\rho_t, \theta_t) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} \Big(\mathcal{L}(y_i, \hat{f}_{t-1}) + \rho\, h(x_i, \theta)\Big).$$

For each specific loss function $\mathcal{L}(y, f)$ and base-learner $h(x, \theta)$, it is difficult to solve the above optimization problem. Friedman proposed to choose a new function $h(x, \theta_t)$ to be most parallel to the negative gradient $g_t(x_i)_{i=1}^{N}$ along the observed data

$$g_t(x) = \mathbb{E}_y \left(\frac{\partial \mathcal{L}(y, f(x))}{\partial f(x)} \middle| x \right)_{f(x) = \hat{f}_{t-1}(x)}.$$

It will help convert difficult optimization problem to below well-studied least-square estimation problem

$$(\rho_t, \theta_t) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} (-g_t(x_i) + \rho\, h(x_i, \theta))^2.$$

C. Sentiment Analysis method

This paper utilizes VADER method for sentiment analysis step. VADER (Valence Aware Dictionary for Sentiment Reasoning), proposed by Hutto and Gilbert [15] is a popular *rule-based* method used for sentiment analysis. The authors construct a sentiment lexicon dictionary, included 9,000 lexicons, each has sentiment score in range -4 to 4, which -4 represents for very negative sentiment and 4 represents for very positive sentiment. For each lexicon, the authors collect its sentiment score of ten people. The score of each lexicon in dictionary is estimated by mean of scores of ten people. This work is done through crowd-sourcing on Amazon Mechanical Turk (AMT)[4]. Basically, VADER receives a single text document as input, evaluates sentiment level of this document and then gives an output score, called *sentiment score*. On each input document, VADER first counts sum of sentiment scores of all words in the document that appear in above sentiment lexicon dictionary. Since each document has different length, the authors proposed a method to normalize sentiment score to range -1 to 1 by the following formula

$$s = \frac{x}{\sqrt{x^2 + \alpha}},$$

where:

- s is normalized sentiment score,
- x is the sum of sentiment score of a text document,
- α is a normalizing parameter, default value is 15.

To improve the expressiveness of sentiment score, these authors introduced five more heuristic rules below:

- **Punctuation**: This will increase magnitude of the intensity. For instance, "*Bitcoin price is increasing!!!*" is stronger than "*Bitcoin price is increasing*".
- **Capitalization**: Especially using ALL-CAPS for sentiment-relevant words will increase magnitude of sentiment and gain more attention from readers. This sentence: "*It is a GREAT NEWS for all traders around the world*" is more expressive than "*It is a great news for all trader around the world*".
- **Degree adverbs**: These words may either increase or decrease sentiment intensity. For example, "*This law is an extremely bad news for Blockchain lovers*" increases the strength of "*bad news*", while "*The Number of*

[4] `https://www.mturk.com`

transactions is slightly lower than yesterday" makes negative contain in sentence less intense.

- **"But"**: This contrastive conjunction may make the clause after "but" dominates the overall opinion. In following example, many people might think second clause is more important: "*More and more researchers work on research problems related cryptocurrency, but very few of them really have innovative insights*".
- **Trigram**: By examining the tri-gram preceding a sentiment-laden lexical feature, the authors caught nearly 90% of cases where negation flips the polarity of the text. A negated sentence would be "*The food here is not really all that great*".

IV. Proposed Method

For predicting Bitcoin movements, suppose that we want to predict the movements at day t, we will derive the features for model based on information at days $t-1, t-2, \dots, t-\ell$, where ℓ is time-lag. Overall, our proposed method includes three main steps: (1) extracting features from social media, (2) extracting features from historical prices, (3) building a classifier using two feature groups above to predict daily up/down trend of Bitcoin. These steps are illustrated by pipeline in the Figure 1.

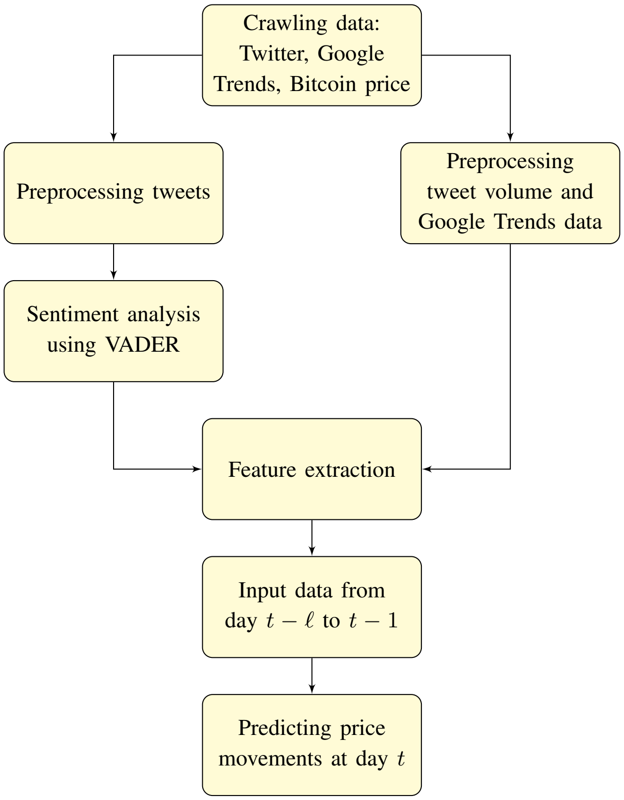
Crawling data: Twitter, Google Trends, Bitcoin price
Preprocessing tweets
Preprocessing tweet volume and Google Trends data
Sentiment analysis using VADER
Feature extraction
Input data from day $t-\ell$ to $t-1$
Predicting price movements at day t

Fig. 1: Prediction pipeline

We now discuss in detail each block in Figure 1.

A. Data preparation

We use Selenium and BeautifulSoup, open-source and easy to use libraries, for crawling dataset from many sources: Twitter, Google Trends and BitInfoCharts.

- Twitter document dataset: This dataset includes *top* tweets which were ranked by Twitter platform about Bitcoin topic. Tweets are collected by following keywords: "bitcoin", "btc", "#bitcoin" and "#btc". We crawled tweets Which were posted during December 1, 2014 to December 25, 2018. For each Tweet, besides tweet contain, we also retrieved other information about tweet creators, the time they've created them, the number of comments, favorites and shares.
- Google Trends dataset: Google provides a set of search trends data. Google Trends website allows users to search for popularity of a particular topic, in this case of Bitcoin, through the Search Volume Index (SVI). Collected data include date and SVI from December 1, 2014 to December 25, 2018.
- Tweet volume dataset: We decide to collect data about the number of tweets created at a given day. BitInfoChart provides data we need from April, 2014 up to now. Collected data includes date, number of tweets generated each day.
- Historical price dataset: We collect historical Bitcoin price from Coinbase[5]. Data was collected from December 12, 2014 to December 25, 2018 including daily **open price**, **high price**, **low price** and **close price**.

After collecting raw datasets, we continue to preprocess them.

i) Preprocessing tweets: Tweet is a type of microblog with a maximum length of 280 characters (formerly 140 characters). Since the length is limited, tweets are often written by users with brief content and accompanied by hashtags (a phrase after the sharp sign). After observing the collected data, we realize that some Twitter users might auto bots created for the purpose of updating Bitcoin information and price. While doing this task, we also find that there are quite a few tweets which only include links to specific articles and newspapers. From these above observation, we propose a method to preprocess tweets:
   - Eliminating tweets created by machines because these tweets do not make sense for sentiment analysis.
   - Filtering out the hashtags of tweets.
   - After two above steps, the filtered dataset will then be tokenized. These tokens will be transferred back into the original form. For examples: "likes" → "like", "worse" → "bad", to make it easy to extract features. This process is also called by another name lemmatization.
   - Removing some strange characters, meaningless in English, in URIs and in URLs.

[5] `https://www.coinbase.com`

- Converting some English abbreviations, such as $'$re, $'$s into the original form.

ii) Google Trends data: We apply the processing method of Erik Johansson [19] to preprocess this dataset. SVI data obtained by searching over 90 days will be divided by week rather than by day. Therefore, to obtain the dataset on a daily basis, we need to do following steps:

- Collecting data by day and then combining them into an array.
- Collecting data by week in the same period of time.
- Adjusting daily data based on weekly data. The adjustment factor is calculated by dividing the weekly SVI by daily SVI.
- Multiplying the SVIs of the remaining days with the adjustment factor.

Table I show an example for this pre-process task.

TABLE I: SVI adjustment.

Date	Daily SVI	Weekly SVI	Adjustment factor	Adjusted SVI
1.1.2014	72	20	0,3	20,0
2.1.2014	96		0,3	26,7
3.1.2014	16		0,3	4,4
4.1.2014	70		0,3	19,4
5.1.2014	61		0,3	16,9
6.1.2014	97		0,3	26,9
7.1.2014	44		0,3	12,2
8.1.2014	32	30	0,9	30,0
9.1.2014	8		0,9	7,5
10.1.2014	13		0,9	12,2
11.1.2014	67		0,9	62,8
12.1.2014	9		0,9	8,4
13.1.2014	63		0,9	59,1
14.1.2014	91		0,9	85,3

B. Social media features

After preprocessing raw dataset, we extract following features:

i) **Daily average sentiment score**: This feature measures average sentiment about Bitcoin of all users represented in tweets each day. We calculate average score instead of directly using result of VADER for the reason that the number of Bitcoin-related tweets is not the same. The score is calculated based method VADER as follow

$$avg\_score_t = \frac{\sum_{d \in D_t} sentiment\_score(d)}{|D_t|},$$

where:

- avg_score_t is the average sentiment score for day t,
- D_t is the set of all tweets about Bitcoin in day t,
- $sentiment\_score(d)$ is tjetotal sentiment score of day d after analysis using VADER.

ii) **Tweet volume**: This feature counts the number of tweets about Bitcoin each day.

iii) **Google Trends SVI**: This feature provides information about the popularity of given search term.

Besides these three features, in this paper, we propose a new feature based on data from Twitter. We call it ***interaction*** feature. The following feature measures how much Twitter users interact with Bitcoin-related tweets each day

$$interact_t = \frac{\sum_{d \in D_t} (likes(d) + retweets(d) + cmts(d))}{|D_t|},$$

where:

- $interact_t$ is total interaction level each day t,
- D_t is all tweets about Bitcoin each day t,
- $likes(d)$, $retweets(d)$ and $cmts(d)$ is number of likes, shares, comments of tweets about Bitcoin each day, respectively.

Although this new social media feature intuitive and simple, it is impact in fact on the results greatly. To our surprise, its value highly correlates with daily Bitcoin price and of course leads to remarkable model performance. We will analyse further the role of this feature in more detail in Subsection V-D while discussing the experimental results.

C. Historical price features

Historical prices plays a vital role in Bitcoin market prediction. Many investors take it as a major source for investment strategies. Our study in this paper is limited to Bitcoin price of Coinbase exchange, one of the largest coin exchanges in the world of cryptocurrency. From Coinbase, we collect information about daily price of Bitcoin and extract following features:

- **Open price**: This is the same as **close price** of the previous trading day. The reason is that Bitcoin is traded 24/7.
- **Close price**: This is the price of a Bitcoin at the end of trading day.
- **High price**: This is the highest price that single Bitcoin achieves on a trading day.
- **Low price**: This is the lowest price that single Bitcoin achieves on a trading day.

V. Experiment and result

A. Dataset

As presented in Section IV, our dataset after preprocessing steps includes two main groups of features collected from December 1, 2014 to December 25, 2018. Since this data has time-series nature, we need to convert it into a form which supervised learning methods can apply. Table II shows an example how we construct features for a data point with time-lag $\ell = 2$.

TABLE II: Features for 2-day time-lag

1-day time-lag	2-day time-lag
pre_1_day_open_price	pre_2_day_open_price
pre_1_day_high_price	pre_2_day_high_price
pre_1_day_low_price	pre_2_day_low_price
pre_1_day_sentiment	pre_2_day_sentiment
pre_1_day_interaction	pre_2_day_interaction
pre_1_day_tweet_volumes	pre_2_day_tweet_volumes
pre_1_day_google_trend	pre_2_day_google_trend
pre_1_day_close_price	

Specifically, each independent variable is a binary value that indicates Up/Down of price of single Bitcoin at day t compared to day $t-1$. Dependent variables of observation at day t are the combination of all features each day in Section IV from day $t-\ell$ to $t-1$ with ℓ time-lag chosen in advance. We note that since the **close price** of one day is equal to the **open price** of next day, **close price** of all days except day $t-1$ will be eliminated from feature space.

B. Evaluation metrics

Our focus is to find an optimal model for predicting Bitcoin price movements. As a by-product, this best model is expected to help investors earn much more money when trading Bitcoin. Therefore, we decide to use two below metrics in our study. The first one is *mean-accuracy* of model when performing $K-$fold cross-validation. From this metric, we hope to gain some understanding of how well the model will generalize when deployed in practice.

$$Mean\text{-}accuracy = \frac{\sum_{k=1}^{K} accuracy_k}{K}.$$

In the above formula, K is the number of folds in $K-$fold cross-validation. Each $accuracy_k$ in formula above is calculated using the next metric.

The second metric used in this paper is test accuracy. This number tests performance of our model on a small unseen data extracted from the original dataset,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

C. Training and testing

We divide our dataset into two subsets: training set and testing set. Training set comprises 90% observations, while testing set includes the rest 10%. We observe that our dataset has very different ranges among features. For example, **daily average sentiment score** has range from -0.012 to 0.293, while **Tweet volume** fluctuates between 7,300 and 155,600. Therefore, we decide to normalize values of features by using $MinMaxScaling$ formula to scale all features to fixed range from 0 to 1. More precisely, for each feature, a $MinMaxScaling$ is typically done via the following formula

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}},$$

where:

- X_{max} is maximum value of concern feature X,
- X_{min} is minimum value of concern feature X,
- X is value before scaling,
- X_{scaled} is scaled value.

Our main experiment is executed as follows: We perform 10-fold cross-validation using Gradient Boosting algorithm on training set for each time-lag case from 1-day to 7-day. The model having highest mean-accuracy on the training set would be the optimal one.

TABLE III: Mean-accuracy on evaluation set each time-lag case

Time-lag	Mean-accuracy	Test accuracy
1 day	54.36%	51.76%
2 days	53.43%	52.38%
3 days	**56.21%**	**57.84%**
4 days	53.14%	46.34%
5 days	52.64%	48.15%
6 days	53.93%	46.25%
7 days	50.79%	46.84%

From Table III, we clearly see that 3-day time-lag is the best fit for our final model. To give readers some ideas about this number, we benchmark our result to other recent papers' results. However, due to the nature of this research area, we need to say that it is extremely difficult to have fair comparisons among these results. For example, the origin of the data used in each study maybe different in many aspects, including social media platforms and time period. Up to now, there is no standard dataset used as a benchmark for evaluating study results in this area. The most clearly related papers to our study is of Kim *et al.* [3] and McNally *et al.* [4]. Using the same 3-day time-lag, our model obtains 57.84%, while the work of Kim *et al.* [3] and McNally *et al.* [4] achive test accuracy 49.47% and 52.78%, respectively. Although we achieve better results than two above papers', no conclusion or interpretation should be made here by the reason of some differences in experiment settings.

D. Impact of social media features on model performance

This session devotes to analysis contribution of information from social media to overall performance of predictive model, especially our new proposed feature. Our first analysis in Table IV is to show correlation between each social media features and Bitcoin price.

TABLE IV: Correlation between social media features and Bitcoin price

Price feature **Social media features**	Bitcoin price
Sentiment	0.45
Interaction	***0.8***
Tweet volume	0.74
SVI	0.79

We clearly see that our new feature outperforms the others in terms of correlation to Bitcoin price. This is a good insight to believe that **interaction** feature also remarkably contributes to our final model. We illustrates these correlation relationships through Figures 2. 3, 4 and 5. The most strangest behaviour when we observe these figures, together with results in Table IV is surprisingly low correlation between daily average sentiment score and daily Bitcoin price. This is somewhat counter-intuitive since many previous studies pointed out that the overall mood about cryptocurrencies from social networks has great impact on Bitcoin value.

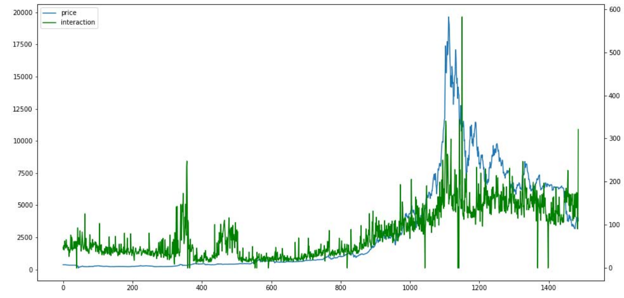

Fig. 2: Correlation graph between interaction feature (green line) and Bitcoin price (blue line)

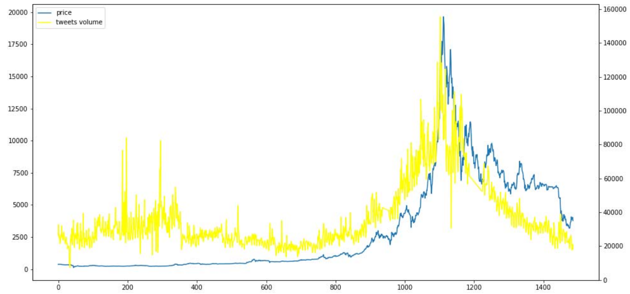

Fig. 3: Correlation graph between tweet volume feature (yellow line) and Bitcoin price (green line)

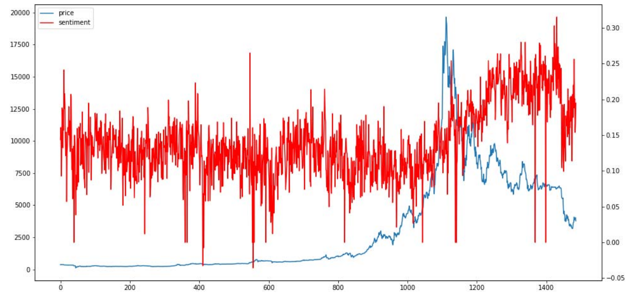

Fig. 4: Correlation graph between average sentiment feature (red line) and Bitcoin price (blue line)

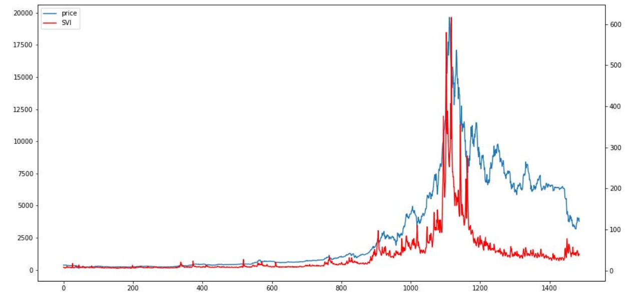

Fig. 5: Correlation graph between Google trends SVI feature (red line) and Bitcoin price (blue line)

In the next analysis, we explain how our novel feature actually improves the model performance. Table V shows how mean-accuracy and accuracy on test data change when we drop social media features extracted from feature space.

TABLE V: Drop features

Features groups	**Mean-accuracy**	**Test accuracy**
All features	56.21%	57.84%
Drop interaction	55.29%	54.22%
Drop interaction + sentiment	53.71%	51.81%
Drop all social media	55%	49.4%

From Table V, we clearly see that using **interaction feature** can improve mean-accuracy from 55.29% to 56.21% and test accuracy from 54.22% to 57.84%. This is a significant improvement for the predictive model when considering that our new feature is easy to obtain and requires low-cost computation.

VI. Conclusion and future work

This paper introduces a novel **interaction** feature derived from social network Twitter resulting better performance for Bitcoin's price movements prediction model. Moreover, our new feature is very intuitive and easy to extract.

We discuss some potential improvement and extension next. A limitation of our current work is that we only collect mood data from Twitter. In the future, we will conduct our study with data obtained from other social media platforms, such as Facebook, Bitcointalk forum[6], etc. We believe that these social platforms, especially Bitcointalk forum would suggest us more relevant Bitcoin-related social signals. Another limitation is the method to evaluate daily sentiment score. Current method using VADER may not be the best choice as we could observe from Table IV and Figure 4. Finally, this study only focuses on Bitcoin. Predicting multiple cryptocurrencies is needed as investors rarely trade only one kind of cryptocurrency.

Acknowledgment

Work presented in this paper is funded by Ho Chi Minh City Department of Science and Technology (HCMC DoST, Vietnam), under grantnumber 12/2018/H-QKHCN.

References

[1] Abraham, J., Higdon, D., Nelson, J., Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), 1.

[2] Mai, F., Shan, Z., Bai, Q., Wang, X., Chiang, R. H. (2018). How does social media impact Bitcoin value? A test of the silent majority hypothesis. Journal of Management Information Systems, 35(1), 19-52.

[3] Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS one, 11(8), e0161197.

[4] McNally, S., Roche, J., Caton, S. (2018, March). Predicting the price of Bitcoin using Machine Learning. In 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP) (pp. 339-343). IEEE.

[5] Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D., Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices. Available at SSRN 2607167.

[6] `https://bitcointalk.org`

[6] Matta, M., Lunesu, I., Marchesi, M. (2015, June). Bitcoin Spread Prediction Using Social and Web Search Media. In UMAP Workshops (pp. 1-10).

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Johansen, E. (2014). Creating daily search volume data from weekly and daily data. Global Payment Trends.

[9] Amjad, M., Shah, D. (2017, February). Trading bitcoin and online time series prediction. In NIPS 2016 Time Series Workshop (pp. 1-15).

[10] Catania, L., Grassi, S., Ravazzolo, F. (2018). Predicting the volatility of cryptocurrency time-series. In Mathematical and Statistical Methods for Actuarial Sciences and Finance (pp. 203-207). Springer, Cham.

[11] Karasu, S., Altan, A., Sara, Z., Haciolu, R. (2018, May). Prediction of Bitcoin prices with machine learning methods using time series data. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[12] Jang, H., Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. Ieee Access, 6, 5427-5437.

[13] bikowski, K. (2016). Application of machine learning algorithms for bitcoin automated trading. In Machine Intelligence and Big Data in Industry (pp. 161-168). Springer, Cham.

[14] Lamon, C., Nielsen, E., Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. SMU Data Sci. Rev, 1(3), 1-22.

[15] Hutto, C. J., Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

[16] Pennebaker, J. W., Francis, M. E., Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.

[17] Bradley, M. M., Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.

[18] Stone, P. J., Dunphy, D. C., Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

[19] E. Johansen, Creating daily search volume data from weekly and daily data, Global Payment Trends, 2014.

[20] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[21] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.