Using Social Media For Bitcoin Day Trading Behavior Prediction

Anna Paula Pawlicka Maule
Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
pawlick5@msu.edu

Kristen Johnson
Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
kristenj@msu.edu

Abstract

This abstract presents preliminary work in the application of natural language processing techniques and social network modeling for the prediction of cryptocurrency trading and investment behavior. Specifically, we are building models to use language and social network behaviors to predict if the tweets of a 24-hour period can be used to buy or sell cryptocurrency to make a profit. In this paper we present our novel task and initial language modeling studies.

1 Introduction

Since 2009, with the introduction of Bitcoin (BTC), cryptocurrency has increasingly gained in popularity. Many investors follow well-known cryptocurrency experts on Twitter and use their advice to guide personal investment strategies[1]. Because people are investing their money and Bitcoin prices can fluctuate quickly, resulting in real life financial gains or losses, models that can rapidly analyze trending discourse on Twitter would benefit investors.

To this end, we are in the process of developing models that use language and behavior extracted from tweets for predicting whether an investor should buy or sell their cryptocurrency. Our overall contributions upon completion of this initial work include a cryptocurrency dataset and models that incorporate both language and social network features for investment action prediction. The ultimate models are intended to comprise a weakly-supervised pipeline for day trading: given tweets from a 24-hour period, predict whether to buy or sell cryptocurrency based off of discussions from that day. This abstract presents our dataset and the initial language-based models of our eventual pipeline.

2 Related Works

The use of social media, specifically Twitter and its social network interactions, to show connections between online discourse and its effects on public opinion has been widely studied in NLP (Sridhar et al., 2015; Hasan and Ng, 2014; Abu-Jbara et al., 2013; Walker et al., 2012; West et al., 2014; Ritter et al., 2010) and a variety of social sciences (Bollen et al., 2011; Burch et al., 2015; Harlow and Johnson, 2011; Meraz and Papacharissi, 2013; Jang and Hart, 2015). There are many works on Twitter sentiment analysis, but closest to our work are those concerning Twitter sentiment and stock market predictions (Kouloumpis et al., 2011; Rao and Srivastava, 2012; Si et al., 2013).

There are relatively few works concerning cryptocurrency analysis and prediction. Of these, a majority use social media sentiment (Jain et al., 2018; Li et al., 2019), volume of tweets (Vidal, 2020), or both (Abraham et al., 2018) as the main feature for prediction. Furthermore, the prediction tasks are typically to predict cryptocurrency prices or whether the prices will rise or fall.

Sentiment is known to be difficult to predict on Twitter. Further, the volume of tweets can be falsely inflated by the actions of bots reporting currency prices, but not contributing to the discourse. Therefore, instead of sentiment or tweet volume, we aim to use the language directly extracted from tweets, their context, and features representing the social network behavior for a buy or sell (investment) prediction.

[1] https://media.consensys.net/i-read-crypto-twitter-for-hours-everyday-here-are-40\\-accounts-that-really-matter-cfecc681379d

3 Datasets

3.1 Twitter Data Collection and Preprocessing

For this work we collected tweets related to cryptocurrency and BTC prices for trading. Rather than collect based on hashtag or keywords alone, we narrowed our research to specific time frames and user accounts. Tweets were scraped from January 2017, when Bitcoin surpassed $1,000 per coin, to its last all time high price in November 2013, until March 2020.

Within these time frames, three types of user accounts were identified for tweet collection to maximize presence of discourse for analysis and minimize tweet noise. These include influential cryptocurrency Twitter accounts, "influencers", which are well known as sources for investment information and thus should provide features for message propagation. Similarly, media accounts from traditional or online news sources are used. Lastly, we include users who frequently tweet about cryptocurrency and have at least ten thousand followers.

Before processing a total of 64,685 tweets were collected. Preprocessing consisted of two main steps. First, we standardized all tweets. This was done by controlling for capitalization, stemming, and removing URLs, character and white space noise, and stop words. Second, we removed irrelevant tweets, i.e., those tweets that do not discuss cryptocurrency trading. This was done by filtering for cryptocurrency-based keywords or hashtags (e.g., Bitcoin, BTC, Ethereum, crypto, cryptocurrency, blockchain, XRP, and altcoin), as well as identifying tweets that correspond to days with an increase or decrease of BTC price. After processing a total of 12,600 filtered tweets were used for experiments.

3.2 Testing Model Dataset

In this preliminary work, we aim to predict whether a user should buy or sell their cryptocurrency shares based on the current day's tweets. In order to create a labeled dataset for training and testing our models, we downloaded the price of Bitcoin from CoinMarketCap [2]. Using this information, we defined a momentum metric as:

$$momentum = \frac{Price_{close} - Price_{open}}{Price_{open}} \tag{1}$$

If the momentum on a given day increases or decreases by five percent on the following day, a script is used to automatically label the tweets as *buy* or *sell*, respectively. Using these labels, we train the models to predict a buy or sell trading behavior. To verify the accuracy of this approach to labeling, a human annotator was asked to label a subset of the tweets. We then compared the annotator labels to our automatically generated labels. The agreement with our automatically generated labels was 84.6%. We used this annotation approach in an attempt to have a fully automatic pipeline, i.e., a user could download tweets from the day, run the model, and get a prediction.

Model	BOW	TF-IDF	DistilBERT
Naive Bayes	66.54%	61.89%	52.89%
Random Forest	78.24%	77.18%	77.12%
Neural Network	50.11%	49.8%	70.17%

Table 1: The columns represent the accuracy of each model when using either a bag-of-words (BOW), TF-IDF vector, or DistilBERT (Sanh et al., 2019) representation of the tweets as features.

4 Experimental Results

This section describes our initial models and features that will serve as baselines for improvement in future models. Table 1 shows the results of using Naive Bayes, Random Forest, and a simple neural network with two dense layers, with either BOW, TF-IDF, or DistilBERT representations of the tweets as features. For the experiment, the data was randomly shuffled and split into 80% training and 20% testing

[2] `https://coinmarketcap.com/`

sets in all the experiments. Interestingly, the Random Forest is more accurate than the hyperparameter-tuned Neural Network models. Further the accuracy of Naive Bayes decreased using DistilBert, but improved significantly the Neural Network. Based on this we are currently exploring a probabilistic graphical model and graphical neural network, in addition to non-language features that incorporate aspects of social network interaction that influences crypto market movements.

5 Future Work and Conclusion

This abstract presents our current work in progress. Predicting day trading behavior, i.e., whether to buy or sell stock, is a complicated task, especially in a volatile asset such as cryptocurrency. Our promising results show that language can be used to successfully model cryptocurrency trading behavior, however there is still room for improvement. We are in the process of adding more informative features extracted from Twitter, including temporal and social information, as well as contextual representations. Furthermore, we are exploring the use of economic framing as a predictive feature. We hope to show that how influential people speak on Twitter affects cryptocurrency trading and that this online discourse can be used for guiding personal investments.

References

Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis. In *SMU Data Science Review: Vol. 1: No. 3, Article 1*.

Amjad Abu-Jbara, Ben King, Mona Diab, and Dragomir Radev. 2013. Identifying opinion subgroups in arabic online discussions. In *Proc. of ACL*.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proc. of AAAI Conference on Weblogs and Social Media*.

Lauren M. Burch, Evan L. Frederick, and Ann Pegoraro. 2015. Kissing in the carnage: An examination of framing on twitter during the vancouver riots. *Journal of Broadcasting & Electronic Media*, 59(3):399–415.

Summer Harlow and Thomas Johnson. 2011. The arab spring— overthrowing the protest paradigm? how the new york times, global voices and twitter covered the egyptian revolution. *International Journal of Communication*, 5(0).

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proc. of EMNLP*.

Arti Jain, Shashank Tripathi, Harsh Dhardwivedi, and Pranav Saxena. 2018. Forecasting price of cryptocurrencies using tweets sentiment analysis.

S. Mo Jang and P. Sol Hart. 2015. Polarized frames on ”climate change” and ”global warming” across countries and states: Evidence from twitter big data. *Global Environmental Change*, 32:11–17, 5.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proc. of AAAI Conference on Weblogs and Social Media*.

Tianyu Ray Li, Anup S. Chamrajnagar, Xander R. Fong, Nicholas R. Rizik, and Feng Fu. 2019. Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics*, 7:98.

Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18(2):138–166.

Tushar Rao and Saket Srivastava. 2012. Analyzing stock market movements using twitter sentiment analysis. In *Proc. of International Conference on Advances in Social Networks Analysis and Mining*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proc. of NAACL*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proc. of 51st Annual Meeting of the Association for Computational Linguistics*.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proc. of ACL*.

Tiago Vidal. 2020. How traders can use twitter to anticipate bitcoin price moves, volume. 02.

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proc. of NAACL*.

Robert West, Hristo S Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *TACL*.