L-Università ta' Malta

Sentiment Analysis to Predict Cryptocurrency Prices

Proposal

ICT3907 - Final Year Project in Computing Science

Supervisors: Dr. Joshua Ellul, Dr. Albert Gatt

Prof.

Jacques Vella Critien - 97500L

13th November 2020

1 Keywords

Cryptocurrencies, Sentiment Analysis, Machine Learning, Data Science

2 Abstract

More than the price, perhaps it's better to be a bit more ambiguous? "Indication of direction and change"?

This research tries to predict the ever-changing cryptocurrency prices by evaluating the sentiment found in social media posts, which directly or indirectly affect these values. Twitter is used as the source of sentiment since it is widely used as a platform where people express their opinion about various topics, including, cryptocurrency price changes, forecasts and other factors which can cause fluctuations in price. Moreover, this relation is tested in several distinct time domains trying to find the optimum interval in which sentiment affects most the price in question. A recurrent neural network model is used for the prediction due to its properties which allow it to base its outcome on features and values from past instances. Finally, this research checks whether using tweet's raw text prove to be a better measure to project cryptocurrency prices than metrics about the same tweets such as polarity, subjectivity and tweet volume, which respectively mean how positive tweets are, how subjective from the author's perspective the tweets are and the amount of tweets posted in the time interval specified.

3 Background

3.1 Problem Definition

Cryptocurrencies' prices fluctuate a lot and this may be the reason why they are not being fully utilised by people on a global scale [1]. Moreover, sentiment expressed in social media affect prices of cryptocurrencies, such as Bitcoin [2]. Out of all the social media networks, Twitter is one platform which can provide enough source of sentiment due to its recent increase in popularity where over 140 million tweets a day are constantly being posted exhibiting opinion, making it a valuable data repository for researchers [3]. The sudden price movements bring about an opportunity for traders and investors. If one has a clear idea of the price trend for the next hour or even day, assumptions can be made when investing which can result in profitable trades, for instance, if it is suggested that the price will surge in the next day, it would be ideal to open a long position [4]. Furthermore, investigation about the correlation between tweets and the price can be performed by assessing in which ways this sentiment best affect price. The accurate prediction of prices is non-linear and hence, not a straightforward problem [5]. The following are some obstacles which were faced in previous studies:

- Twitter may include duplicates tweets which are tweeted by bots or for advertising purposes [6].
- Not many pre-classified or annotated tweets were available to researchers to perform sentiment analysis using a supervised learning approach [3].
- Tweets may contain noise which makes it even more difficult for the tweet to be given an accurate polarity score [7].
- Moreover, around half of the tweets are of a neutral sentiment, hence, they do not quite help in determining the trend of the price [8].
- Even worse, tweets' sentiment usually stays positive irrelevant of possible price drops during the same time. This happens because users who tweet might be still expressing interest in other properties of the currency [8].
- Some tweets may be of a sarcastic nature, being more difficult to correctly identify their polarity [9].
- Using regression architectures instead of specialised models such as Long Short-term Memory Networks which are able to remember certain market behaviours which might appear in historical data [6].

3.2 Personal Motivation

First of all, as a student following a degree in Computer Science, I feel an automatic urge and motivation towards contributing to the evergrowing field of cryptocurrencies and digital currencies. In addition, I am inclined towards the Fintech [10] industry and have always been interested in trading. As previously explained, cryptocurrencies'

prices fluctuate and hence, if one is given indications which are proven to be accurate about potential price changes, this research could prove to be a tool for investors. Apart from that, if the correlation between Twitter tweets' sentiment and price changes exists and is accurate, this research could push more users to engage in expressing their sentiment, making this predictor even more precise.

4 What has been done

4.1 Reading

During the summer months, my supervisors suggested that it would be a good idea to start searching and reading resources related to the topics which are tackled in the research. Consequently, the following topics were covered in my readings:

- **Sentiment Classification**: [11] was the main source which provided knowledge about this concept. This resource contained detailed explanations ranging from the most basic terms, such as recall, precision and F-measure, to multiple techniques such as handling negated verbs, Naive Bayes Classification, using N-Grams Language Models, representations of words and vectors and Embeddings. Moreover, this book elucidates the notions of neural networks and their possible implementations to classify sentiment.
- **Sentiment Analysis in Twitter**: [9,12–15] are outcomes from the SemEval [16] workshops and these include the tasks attempted. The tasks in question include:
  1. Classifying tweets as being positive, negative or neutral [12].
  2. Classifying out-of-domain and sarcastic tweets as being positive, negative or neutral [9].
  3. Determining polarity towards a topic from a set of tweets [13].
  4. Classifying tweets in a five point scale, that constitutes of classes for being highly positive, positive, neutral, negative and highly negative [14].
  5. Handling multiple languages and using information about the tweets' authors [15].
- **Predicting Prices using Sentiment Analysis**: Several papers which deal with trying to predict prices using sentiment analysis were also read.
  1. [2] shows how a recurrent neural network is used to try and predict future prices by taking into consideration closing prices and output from a sentiment analyser which examines 2585 manually labelled tweets.
  2. [3] tries to predict Microsoft's stock price movement by sentiment analysis. It uses word2vec representation of 3126 human annotated tweets as features in a machine learning model in order to analyse the sentiment in the other tweets. The stock's price trend is labeled using a simple program which lists the current day with a value of 1 if the current day's price is bigger than the previous day's or a 0 if it is smaller. The best results were achieved using a 3 day window and the classifier was trained using a logistic regression algorithm.
  3. [6] compares the performance of using a Multi Layer Perceptron, a Support Vector Machine and a Rain Forest Algorithm to try and predict the daily market movements of Bitcoin, Ethereum, Ripple and Litecoin using market data, twitter data or the two combined. This research suggests that Multi Layer Perceptrons using both market and twitter data perform best in most cases. However, suggests that LSTM cells should be tested in future works.
  4. [8] uses a linear model to predict Bitcoin price when inputted tweets and Google Trend data. This research ignores tweets' sentiment as it shows that apart from half of the tweets being neutral, the ones which actually had a side, involved positive sentiment even in times when price was falling. On the other hand, it shows that the volume of tweets actually correlates with the price. In fact, as future work, this research asks whether this relation holds for varying pricing environments.
  5. [7] uses a lexicon-based approach to label tweets, hence, being unsupervised. However, it concludes that this might be the reason why the study's results were not better and suggests building a labelled data set to perform supervised techniques.
- **Tutorials on how to implement neural networks using TensorFlow**: [17] contains documentation and tutorial on how to use TensorFlow and Keras to build a neural network.

4.2 Data collection

for opening quotes use ` and for closing '

1. **Tweets Dataset**: As a dataset for tweets, this research will use the 'Bitcoin tweets - 16m tweets' [18] dataset from Kaggle which contains 16 million tweets from every minute ranging from 1st January 2016 till 29th March 2019. This dataset contains the author's username, the author's full name, the timestamp, the actual text, the url and the number of likes, retweets and replies.

2. **Bitcoin Prices Dataset**: In order to obtain Bitcoin's prices, I wrote a Python script which is able to obtain historical cryptocurrency prices for every minute ranging between two dates. This data includes the opening, closing, highest and lowest prices and timestamp of every minute.

In addition, another Python script is written which is able to combine the two aforementioned scripts into one by joining the every tweet to the corresponding Bitcoin prices at that timestamp. Finally another python script was prepared which is able to get the average metrics of the tweets according to the parameterised time frequency. For instance, if an hour time frequency is passed, a new dataset will be created consisting of the opening, closing, highest and lowest prices of Bitcoin together with number of tweets in that hour and the average polarity and subjectivity, which are obtained using an unsupervised lexicon-based approach, of the tweets in that hour.

4.3 Model creation and testing

A basic model was created to be able to predict one step in the future when passed a daily or an hourly dataset prepared from the data collection and cleaning methods explained above. This is a recurrent neural network model making use of LSTM cells. Despite already showing positive signs of prediction, this is still in early stages and as will be explained below, this requires more testing and variations to obtain better trend accuracy results and confirmation of the complete elimination of possible overfitting.

5 What has to be done

5.1 Way forward

- Start manually labelling some tweets for supervised machine learning for classifying tweets.
- Perform data cleaning on the tweets to remove noise such as URLs, duplicates or punctuation marks..
- Continue optimising the current model to try and bring out better results.
- Try using VADER for unsupervised sentiment analysis instead of TextBlob to compare performance of the two.
- Change the current model to predict multiple time steps in the future.
- Create a new model which takes in tweets raw text instead of tweets' averages to compare the performance between the two.

5.2 Research Questions

- Which is the best suited model to classify sentiment in twitter tweets?
- Does unsupervised lexicon based approaches perform better than supervised machine learning when classifying tweets?
- Is it better to pass in the model tweets as raw text or averaged properties about the tweets in a particular time frame?
- Which time window produces the most accurate price trends from twitter sentiment?
- How many time steps in the future is the model able to predict?

5.3 Hypothesis

Finally, this research should be able to answer the hypothesis question of whether **Cryptocurrency price trend can be correctly predicted from Twitter tweets**.

5.4 Research Outcome

Once the study is finalised, this research could be used as a guideline by investors to have an insight of where the price of the cryptocurrency is heading, hence, making more smarter decisions with lower risks. Moreover, if the results that twitter sentiment actually can predict a cryptocurrency price, it could stimulate more users to engage in expressing their opinion,. Thus, apart from creating more input for the classifier, future researchers will have even more data at their disposal for training their models. Furthermore, from the answers of the research questions, this research could help future researchers to choose more easily the tools needed when implementing their model and eliminate some doubts about any implementations outperfoming others.

References

[1] Y. B. Kim, J. Kim, W. Kim, J. Im, T. Kim, S. Kang, and C.-H. Kim, "Predicting fluctuations in cryptocurrency transactions based on user comments and replies," *PLOS ONE*, vol. 11, p. e0161197, 08 2016.

[2] D. Pant, P. Neupane, A. Poudel, A. Pokhrel, and B. Lama, "Recurrent neural network based bitcoin price prediction by twitter sentiment analysis," pp. 128–132, 10 2018.

[3] S. Pagolu, K. Challa, G. Panda, and B. Majhi, "Sentiment analysis of twitter data for predicting stock market movements," 10 2016.

[4] J. Chen, "Long positions." `https://www.investopedia.com/terms/l/long.asp`, October 2020.

[5] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural network," vol. I, pp. 1 – 6 vol.1, 07 1990.

[6] F. Valencia, A. Gómez-Espinosa, and B. Valdes, "Price movement prediction of cryptocurrencies using sentiment analysis and machine learning," *Entropy*, vol. 21, pp. 1–12, 06 2019.

[7] O. Kraaijeveld and J. De Smedt, "The predictive power of public twitter sentiment for forecasting cryptocurrency prices," *Journal of International Financial Markets, Institutions and Money*, vol. 65, p. 101188, 03 2020.

[8] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," 2018.

[9] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov, "Semeval-2014 task 9: Sentiment analysis in twitter," 2014.

[10] J. Kagan, "Financial technology – fintech," Aug 2020.

[11] D. Jurafsky and J. Martin, *Speech and Language Processing*. 01 2000.

[12] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson, "Semeval-2013 task 2: Sentiment analysis in twitter," 2013.

[13] S. Rosenthal, S. Mohammad, P. Nakov, A. Ritter, S. Kiritchenko, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter," 2015.

[14] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," 2016.

[15] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," 2017.

[16] "What is semeval?." `https://semeval.github.io/`.

[17] "Tensorflow tutorials." `https://www.tensorflow.org/tutorials`.

[18] "Bitcoin tweets - 16m tweets." `https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329`.