|  |  |
| --- | --- |
| **Project Proposal for PLP Certificate** | |
| **Project Title** | Prediction of Bitcoin prices using Sentiment Analysis, and Topic Modelling from Social Media Forums |
| **Project Members** | Gui Lin (A0215294E), Song Bingheng (A0215496X)  Teo Wen Lin (A0035184J),  Wang Sixiang (A0215475A), Xiong Hui (A0215431U) |
| **Overview** | Literature Review  Previous research has shown that the fluctuation in bitcoin prices have some correlation with various factors, one of which is the value placed on it by end users. Researchers have used several methods to determine whether and the extent to which user sentiment can explain the variation in bitcoin prices.  Kaminski and Gloor (2014) [1] analysed Twitter data over 104 days and compared this to the fluctuation in bitcoin prices, using a small lexicon of 15 words describing the polarity of financial sentiment. They found that the Twitter sentiment only mirrors and does not predict bitcoin prices.  Karalevicius et al (2017)[2]identified an interaction between media sentiment mined from bitcoin-related news portals and the price of Bitcoin, and that there was a tendency for investors to overreact on news in a short period of time. In their research, the finance-oriented sentiment dictionary created by Loughran and McDonald (2011)[3]was used to identify positive and negative sentiments in the news articles.  Linardatos and Kotsiantis (2018)[4] used several deep learning models, using Google trends data and bitcoin-related tweets over 2 years (2017 to 2018) as input to predict bitcoin price for the next day, with error rates between 0.99 and 2.66%.  This project  In this project, we aim to build similar models to predict bitcoin price for the next day, using a sentiment analysis of social media comments on 2 popular bitcoin forums (<https://bitcointalk.org/> and <https://www.reddit.com/r/BitcoinMarkets/>)  over 4 years from 2017 to 2020, as well as labelled financial sentiment text data from Financial Phrase Bank and Financial Tweets. We also aim to extract the main topics discussed on these forums using topic modelling methods. |
| **Design** | **We have 2 tasks, the first task is sentiment analysis about bitcoin price. The second task is Topic Modelling involving bitcoin price**   1. **Sentiment analysis about bitcoin price:**   1(a) Sentiment score about bitcoin comments   1. Collect financial labelled data 2. Collect bitcoin text comments, and manually label the sentiment of a few bitcoin text comments 3. Deep learning model training (fine-tuning BERT for sequence classification) to get the sentiment score for each bitcoin comment   1(b) Prediction of bitcoin price   1. Collect bitcoin prices from 2017 to 2020 2. Collect Google Trends data on ‘bitcoin’ from 2017 to 2020 3. Deep learning model (LSTM) to predict each day’s closing bitcoin prices based on sentiment scores of the previous x days, number of comments in the previous x days, Google Trends values, and past x days’ bitcoin price (where x is a value between 1 and 5)   For task 1(a):  Model inputs (financial labelled data):  Financial Phrase Bank (from financial news articles), created in 2014: <https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10>  Financial Tweets with Sentiment, created in 2018/19: <https://www.kaggle.com/vivekrathi055/sentiment-analysis-on-financial-tweets?select=tweet_sentiment.csv>  Model output:  Predicted sentiment scores  Deep learning model:  Fine-tuning BERT  For the top layer, use softmax activation function to output the probability as a representation of the sentiment score.  For task 1(b):  Model inputs:   1. Average sentiment score based on the prediction of the deep learning model, grouped by day   (2)Total number of comments from both forums, per day  (3)Google trends on bitcoin, per day  (4)Each day’s bitcoin closing price from 2017 to 2020  Model outcomes:  (1)The difference between closing price of the current day and next day  (2)The fluctuation in price for the next day(high –low)  (3)The volume of trades for the next day  **2. Topic Modelling**  The second task is to identify the top few topics discussed by bitcoin users every month. This exercise has business value because it is a relevant factor to the bitcoin price. Topic modelling will be done using the text data extracted from bitcointalk.org and reddit.com/r/BitcoinMarkets, using the following unsupervised methods:  (1)K-means Clustering  (2)Latent Dirichlet Allocation |
| **Scope of Work** | Dataset sources (crawl using python requests library):  <https://bitcointalk.org/>  <https://www.reddit.com/r/BitcoinMarkets/>  If time permits, to crawl the following forums as well:  <https://bitcoin.org/en/community>  <https://bitcoin.stackexchange.com/>  Google trends data:  <https://trends.google.com/trends/explore?date=2017-01-01%202017-09-01&q=%2Fm%2F05p0rrx>  Bitcoin historical prices obtained from:  <https://finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD>  Data cleaning and processing: converting emojis and emoticons to words, remove irrelevant symbols, url links, outliers and missing values, dupilcates  Model training for prediction of trends: LSTM / BERT  (If time permits) Uploading of model onto public cloud service (e.g. AWS) for prediction of future prices |
| **Effort Estimates** | Estimated time taken for each task (with overlap)  Data crawling: 10 days  Data cleaning and pre-processing: 10 days  Model training, testing and re-training: 8 days  Uploading model to cloud service and testing: 6 days |
| **References** | [1] Kaminski, J. and Gloor, P.A. (2014), “Nowcasting the bitcoin market with twitter signals”, CoRR abs/1406.7577, available at: <http://arxiv.org/abs/1406.7577>  [2] Karalevicius, V.,Degrande,N. andDe Weert, J.“Using sentiment analysis to predict interday Bitcoinprice movements”.The Journal of Risk Finance Vol. 19 No. 1, 56-75,Emerald Publishing Limited, 2018  [3] Loughran, T.and McDonald, B.(2011), “When is a liability not a liability? Textual analysis,dictionaries, and 10-ks”, The Journal of Finance, Vol. 66 No. 1, pp. 35-65, available at: <http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x>  [4] Linardatos, P.and Kotsiantis,S.“‘Bitcoin Price Prediction Combining Data and Text Mining” in Advances in Integrations of Intelligent Methods, Chapter 3, Springer, 2018 |