**Raditya Fahritama – rkf5230@psu.edu**

**Scientific Review**

**Aim of the Research**

**Heung Seok Joon et al**. presenting some technical approaches and model to Predict future Bitcoin’s price by developing model with bitcoin price history and Twitter sentiment. Another focus of this research is to find a partial correlation between the price fluctuation of Bitcoin and the fluctuation of the sentiment classes using a machine learning algorithm. The author feels that analyzing tweets using Twitter sentiment analysis is a significant step to predict the price movement.

**Georgoula et al.** tried to know what are the factors that will affect the price of Bitcoin in short-run and long-run. Focus of the paper is to study the dynamics governing the formation of Bitcoin prices by focusing on Twitter sentiment as an explanatory factor along with other economic and technological variables. And constructing sentiment ratio for Twitter user on daily basis. This ratio is based on notion that standard economic theory might not adequately describe changes in Bitcoin prices, and one should also take short-run speculative investment incentives or expectations into account. The authors use Google trends and Wikipedia as a proxy for public sentiment and finds a positive correlation between these measures and the price of Bitcoins.

**Pagolu et al**. feel that the thriving culture of social media make the information about public feelings becoming abundant. The authors feel that social media has an impactful influence on overall public opinion. The research focused on forecasting stock prices by using Twitter sentiment and developing analyzer that can judge the type of sentiment that present in a single tweet. The authors also classified the tweet into three categories: Positive, Negative, and Neutral. The authors also want to identify if there is a correlation between stock prices movement to public emotions about the company that is expressed through tweet.

**Scientific Challenge**

***The Data Input is raw data, resulting in noisy dataset***

Often, the arbitrary nature of a tweet’s structure will be a special challenge for researchers to work on. The casual fashion of people when tweeting can’t be predicted by any means. Also, tweets have some numerous exclusive features. These features such as retweet, emoticon sharing, image sharing, location sharing, user references, and other, these unique features can be a hindrance when working on the data that is based on a tweet that has those features. Thus, the input has to be extracted correctly.

***Determining which Preprocessing steps that are compatible for the model***

The rawness of input data from twitter has to be structured in order for the researcher to work. So that the action of standardizing the input data is important. Researcher has to know how to preprocess raw data that is applicable and can be easily learned by different classifiers. Eventually, require extensive number of preprocessing steps for the input data to be understandable by the classifier. Determining which preprocess steps are right for some model is not an easy task. This process can easily be the main challenge of the researcher.

***Treating Missing Values***

When working with data, we cannot expect that the input data is perfect. Most of the time, there are always flaws in them. One of the most common problems in data preprocessing is dealing with missing values. In order to achieve a stellar result, we have to treat the missing values right. **Pagolu et al.** dealt with this challenge by some technical approach that is concave function ‘(y+x)/2’ to approximate the missing values and eventually, filling all the gaps.

**Contribution and Difference from Related works**

Other works predict the sentiment with different model, different algorithms, and from different sources. **Heung Seok Joon et al**. used Random Forest Regression algorithms with various tweets from twitter as the input data. The author did some general preprocessing steps like converting and replacing character. And the author used an analyzing technique called VADER (Valence Aware Dictionary and Sentiment Reasoner). The sentiment feature of VADER returns a polarity score for a compound. The polarity score for VADER is between -1 and 1, wherefrom -1 to 0 is negative, 0 is neutral and 0 to 1 is positive.

**Georgoula** et al. conducted study by using Twitter Sentiment Analysis, Support Vector Machine, and some models of regression. The author used some extra methodology that is Corpus Vectorization. Using mathematical representation via the employment of Vector Space Model. The motivation is to transform each document into vector that contains the word that belongs to the document and the frequency. The author also featured econometric methodology, that approach follows a number of steps associated with the analysis of time series data.

**Pagolu** et al. used social media microblogging, Twitter, to forecast stock price, it accurately reflects the public's cognitive condition and thoughts about current events. The application of the author’s research is the sentiment analysis and the supervised principle of machine learning of extracted tweets. It also uses feature extraction that has two methods: N-gram Representation, that is a full corpus of related text is parsed which are tweets in the present work, and every appearing word sequence of length n is extracted from the tweets to form a dictionary of words and phrases, and Word2Vec Representation, that are functions by mapping words to 300-dimensional vector representations.

**Strength and Weakness of the research**

One of the notable weaknesses of the research is only using one algorithm. This resulting in a limited point of view on how other machine learning algorithms perform in the model. Either the results will be poorer or greater, considering using at least 3 algorithms will be more challenging. Lacking variety in interpretation result is considered a deciding factor. And other notable weakness is not considering other aspects of the explanatory factor as the determining variable. Like Hashtags, Emoticon, sentiment of the user, etc. Judging by only the words produce one-dimensional result. Developing model with these factors will bear more sophisticated result.

Strong point of this research is that it used all kind of tweets. Not just tweets that are related to Bitcoin or Cryptocurrencies only. This is a brilliant decision because the variety of result will be broader and resulting in variety of deciding factors of price fluctuation. And the research categorized the sentiment into categories. Making the final result is more interpretable by non-technical readers or stakeholders. The research also covers multiple technical approaches and hefty number of strategies on dealing with the problem.

**Relevancy and Lessons Learned**

The research focused on how to read Bitcoin’s price volatility by analyzing sentiment from social media as the explanatory factor. The research covers issues that is similar for our proposed study. Not only by technical means, but also from methodology perspective. In our case, which is analyzing sentiment from reddit, shares similarity with this research. Like predicting prices with sentiment categories. And identifying how the sentiment correlate with prices. These papers have all of them. We believe that we now have clear future plan on how to do the project, and what to expect for the challenges when doing this project, with these papers as guidance.

**Reference**

Sattarov, Otabek, Seok Jeon, Heung, Oh, Ryumduck, Dong Lee, Jung. “Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis.” 2020.

Georgoula, Ifigeneia, Pournarakis, Demitrios, Bilanakos, Christos, Sotiropoulos, Dionisios N., Giaglis, George M. "Using time-series and sentiment analysis to detect the determinants of bitcoin prices." 2015.

Sasank Pagolu, Venkata, Nayan Reddy Challa, Kamal, Panda, Ganapati. “Sentiment Analysis of Twitter Data for Predicting Stock Market Movements.” 2016.