Predicting Doge Coin Price Bubbles using Epidemic Modeling

Garimendra Verma
Georgia Tech / Atlanta, GA
gverma30@gatech.edu

Surya Teja Adluri
Georgia Tech / Atlanta, GA
sadluri3@gatech.edu

1 Introduction

As the title suggests, our main goal is to predict price bubbles of doge coin, a type of crypto currency, using epidemiological models such as HMM, which previously used for Influenza surveillance detection. We are also going to incorporate social media data, and it's sentiment analysis to better predict the outcome. Previously, researchers used only social media data to predict crypto currency prices, but the results proved insufficient to make profits. The reason behind this is the volatility and price bubbles of the crypto market. According to Investopedia, a bubble is an economic cycle that is characterized by the rapid escalation of market value, particularly in the price of assets. This fast inflation is followed by a quick decrease in value, or a contraction, that is sometimes referred to as a "crash" or a "bubble burst". [(Le Strat Y, 1999) A price bubble comprises five phases in the following order: 1. Displacement 2. Boom 3. Euphoria 4. Profit 5. Panic.

To efficiently predict the behaviour of the doge coin, we need a model that not only predicts the price movement according to the sentiment analysis, we also want a model that detects the boom and panic phases of the crypto price bubble. This essentially means we need a model that can predict something as early as possible, to achieve this purpose epidemiological models can be utilised. One advantage of using an epidemiological model for predicting the doge-coin behaviour is, we can use wide number of tools and techniques to explain the trend behaviour and also to predict the outcomes. For instance, surveillance of an infectious disease is a quite well-studied subject, current surveillance systems try to predict the epidemic as early as possible, because it helps in prompt intervention, which is very important due to the threat of new infections, as well as emergence of new virus strains. As discussed above, this property of Epidemiology models of such early detection can be useful in detecting the boom and bust phases of the crypto bubble. Hence, in this project we are predicting the doge coin's price using such epidemiological models. Nevertheless, epidemiological signals alone can't guarantee the accurate price prediction, so we are also using social media data.
Recently in April/May, doge coin's sudden rally in the crypto market caught everyone's eye. News and articles have been published on how tweets by Tesla CEO have been instrumental in the movement of the Crypto market, causing significant price fluctuations, which showed doge coin is susceptible to the changing winds of the social media sentiment. In this project, we collected data from twitter using Twint API, we crawled the entire twitter data base in the doge coin's rally time frame to extract tweets related to particular hashtags which were trending and contributed to the rise of doge coin price. Then, we performed a lot of data cleaning, and data pre-processing to transform the data, then we performed sentiment analysis for every tweet, and transformed the scores into an time-series data. Later, we use HMM model and train it on social media analysis to predict the distribution of probability of that data point being in an epidemic state, detailed methodology of how we are going to accomplish this is written in the below sections. After that we will also study the effect of influenza data and covid US new infections data on doge coins' market using HMM.

2 Related Work

Following the search results, the Sentiment Analysis has been studied in abundance for various tasks, often including social media data from identifying hate speech over the internet, to forecasting Crypto currency prices from Tweets [A. Jain and Saxenae (2018)]. By assigning quantitative values statements that are subjective in nature, Sentiment analysis acts as a form of textual analysis. There exists a herding nature, and tends to be very strong for currencies which share a small market cap. This is different from the scenarios where particularly famous personalities attach themselves to a coin, and their online activity becomes correlated as key driver of the prices in the market. Bitcoin price fluctuations have been studied using twitter sentiment analysis[I. Georgoula (2015)]. They built various regression models, and a Support Vector Machine (SVM) for a 89.6 percent accuracy, finding only a short-term correlation between the Bitcoin price and the positive twitter sentiment. Another approach for this problem is to use VAR or Vector Auto regressive Model and Granger - causality as performed by Garcia and Schweitzer to test a lexicon-based approach to find that changes in Twitter sentiment polarity precede Bitcoin price fluctuations. As across exchanges, the crypto currency prices vary significantly, rendering the result unclear.

NLP and Deep learning techniques have been used for the prediction of stock prices. Abdullah et al [(Saeed Abdullah, 2011)] presented how news spreading dynamics on Twitter can be studied using the Epidemics models like SIR, and based on the work of HMM on influenza epidemics [(M. A. Martinez-Beneito and Lopez-Maside, 2008)]. In work by Kuhle [(Kuhle, 2018)], the spread of investment ideas is studied using the SIR model, which governs the rate at which investors acquire and abandon their views on particular assets. The paper differentiates two scenarios: investors are infected after buying an asset and are cured when they sell it. The other scenario has rational expectations formed by agents, influencing them to wait for further appreciation of overpriced assets. The search queries from Google, i.e., the data from Google Ngrams and Google trends, are used as a proxy for the mass of infected agents, resulting in the same fever curve pattern for the asset prices and the queries. According to the prediction by their approach, which models booms and busts in asset prices as an epidemic process, the prices peak earlier than the mass of infected agents does, i.e., prices peak earlier than the search queries/literature mentions them.

Since there is a fundamental similarity between how the news spreads on Twitter and how the infections spread, one can successfully use Epidemiology models on Twitter data. In this paper, they showed the advantages of using an already developed epidemiology model. For example, we can show that trend dynamics on Twitter can be detected by extending a surveillance system used to detect influenza epidemics. The Bubble-like behavior has been studied using social media and epidemic modeling data in a study by Gorse and Phillips [(Phillips and Gorse, 2017)] . It is proposed that the early stages of the formation of a bubble can be predicted by studying the patterns in social media usage. The detection of epidemic and non-epidemic stages of trading volume and social media usage is performed using Hidden Markov Models too. The model shows false positives for bubble detection, even though it profits from the dynamic sliding window approach, which is implemented to ensure that the model is always considering the most recent social media and the trading volume data as it becomes available.The paper considers social media indicators such as posts, subscriber growth, and new authors, but it does not rank the trending posts by accounting for their upvotes, downvotes, or re-shares, which is a valuable metric for relating the social media usage to trading volumes.

There is a need to find a more compelling medium that includes significant directed activity by potential investors rather than the case of Google search activity. The paper titled "Twitter mood predicts the stock market" [(Johan Bollen, 2011)] provided vital insights on how different moods in the Twitter data influence the stock market. They analyzed the Twitter feeds by OpinionFinder, which measures positive vs. negative attitude, and Google-Profile of the Mood States, which measures mood in terms of 6 dimensions. Daily variations in public mood states show a statistically significant correlation to daily changes in Dow Jones Industrial Average(DJIA) closing values. Some dimensions like Calm increased the Neural Network model's accuracy in predicting up and down changes in DJIA closing values to 87.6 %. This work reinforced our belief in the correlation between social network activity, especially Twitter, and the prediction of stock prices.

3 Proposed Method

3.1 Intuition

Previously, researchers used only social media data as an indicator to predict cryptocurrency prices(I. Georgoula, 2015), but the results proved insufficient to make the profits. The reason behind this is the volatility and price bubbles of the crypto market. There has also been some research on how cryptocurrency price bubbles have previously been linked with the epidemic-like spread of an investment idea (E. Shtatland, 2008). The idea is similar to how epidemic diseases spread when infected by a virus; a cryptocurrency market may boom or burst when infected with an investment idea. In predicting such epidemic diseases, especially influenza monitoring, researchers utilized Hidden Markov Model to detect the pandemic in an early stage (M. A. Martinez-Beneito and Lopez-Maside, 2008).

The intuition behind this project is to study the application of previous works from epidemiology on different domains other than Epidemiology. This project aims to extend the idea of the hidden Markov model's usage in the early detection of Influenza disease to the cryptocurrency market. We chose the cryptocurrency market because there is a direct correlation as to how the price of a crypto market booms/bursts similar to an epidemic disease. The novelty of our idea is we want to show how previous research work of unique mathematical models in Epidemiology can easily extend to different domains. We also wish this project gives enough motivation to utilize the previous work done in epidemiology in various fields other than Epidemiology. A detailed approach and further practical details about the project are discussed in the next section.

3.2 Description

3.2.1 Approach

As mentioned in the previous section, the crypto market is prone to sudden changes due to new investment ideas in the market. In order to best study this behaviour we attempted to predict doge coin price bubbles. As explained in the previous sections, crypto price bubble comprises of five phases in the following order: 1. Displacement 2. Boom 3. Euphoria 4.Profit 5. Panic. To efficiently predict the doge coin bubbles, we need a model that detects boom, euphoria and panic phases in the market as early as possible. To achieve this we are extending the work of (M. A. Martinez-Beneito and Lopez-Maside, 2008) in early detection of influenza pandemic to doge coin. We are using an HMM model that is previously been used on influenza data to successfully detect early stages of doge coin market.

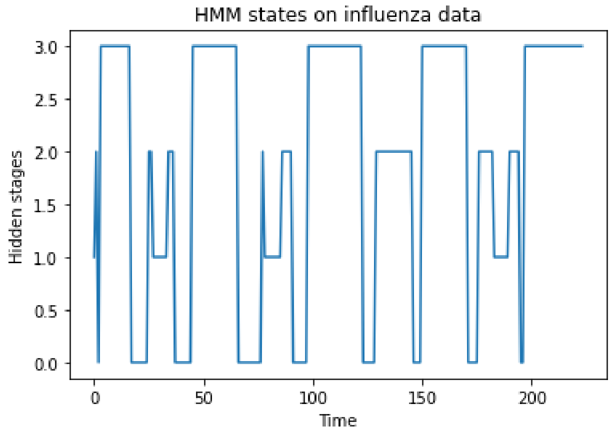

Figure 1: Influenza data after trained using HMM

As seen from the above image when an influenza time series data is trained on Hidden Markov model, the algorithm can successfully detect hidden states in between low-epidemic and high-epidemic states. Researchers used those hidden states to detect the disease even before happening. This similar approach of early detection of influenza is used in this project to detect early rise in the price of doge coin. First, we tried to predict doge coin price bubbles using epidemic modelling, later we are tried to find the effect of influenza and covid data on the price of doge coin.

3.2.2 Data Collection

First, we collected twitter data on doge coin, as we used sentiment data extracted from twitter as a source for identifying investment ideas that may trigger crypto market. The first step for collecting the desired tweets involved finding the hashtags for Dogecoin such as 'tothemoon', 'Musk', 'dogecoinrise', 'dogefather', 'doge', etc. and using the TWINT library which is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from profiles without using Twitter's API to obtain the initial raw data in English language from twitter based on such hashtags and keywords. The data obtained in a python dataframe is saved as a CSV with attributes including the user ID, a unique identifier, the time stamp, and how many times the tweet was "retweeted".

Being a very volatile market, and very recently been proven to be driven by the updates on online social media platforms like Reddit, Twitter, etc. the crypto currency prices show a trend of boom and burst cycles. In order to study the epidemic behavior exhibited here, we collected the Doge Coin prices around the timeframe when Elon Musk claimed via Twitter that “SpaceX is going to put a literal Doge coin on the literal moon”. Such events which often started as memes on the Internet have shown social media’s direct effect on the crypto markets, e.g. the short squeeze of Game Stop in January. The Doge Coins saw a sudden rally in the Cryptocurrency market in April-May, hence we use the data from CoinBase for the period from March 15, 2021 to August 15, 2021. This includes the Open-High-Low-Close (OHLC) prices and volume of the coin traded for each day in USD.

We collected the Influenza data from a website called data.world posted by the California health and human science Link. We pre-processed this data by filtering out the other types of infections and only focusing on the total infections, we also grouped the data by region wise and calculated the sum of all regions. This data is collected for years from 2010-2020. The collected data is plotted and shown in the figure 4. From the figure we can see the influenza data is similar to the boom and bust cycles of the doge coin prices. Also, we can observe once the infections goes to epidemic stage, the infections begin to rise continuously, and when in the non-epidemic stage there is no movement in the number of infections. From the work of [(M. A. Martinez-Beneito and Lopez-Maside, 2008)], using HMM we can model the data to predict the early detection of that epidemic stage. This early detection is used in this project to predict the bubbles of doge coin.

3.3 Model

The main idea behind using a HMM model in this project is to determine epidemic and non-epidemic phases using a five-stage Markov model. In this project HMM model that is previously trained on influenza epidemic will be used to detect epidemic and non-epidemic states of sentiment analysis. The hidden states are unobserved, let’s say E_t is an unobserved random variable to denote whether the system is in the epidemic state (1) or not(0) at time t.

Emission Probabilities: The hidden states have emission probabilities, they give the likelihood of seeing particular output values. In this project, we sampled emission probabilities from differenced time series data or gaussian noise. I_t represents the difference between the time series at t and t-1, I_t is sampled from the auto regressive process of order 1 for the epidemic state, or sampled from the gaussian noise distribution for the non-epidemic state. Which means in layman terms, epidemic state has interrelated changes, where as the non-epidemic state has random but very small changes. The idea of using auto regressive models is appropriate to model time series data dynamics during financial asset bubbles, is taken from the study[E. Shtatland (2008)].

Transition Probabilities: HMM algorithm transitions among the hidden states by following a transition probability matrix. These transitions also assume a markovian property, to achieve this later we used moving window technique to feed the data to the HMM model.

$$P_{k,\ l} = P(E_{t+1} = l,\ E_t\ =\ k)$$

Once the model and probabilities have been created, using expectation maximization (EM) estimates of optimal parameter values can be found.

4 Experiments/Results

4.1 Testbed

In this project, we aim to predict the price bubbles of doge coin, This is our overall Experiment Structure.

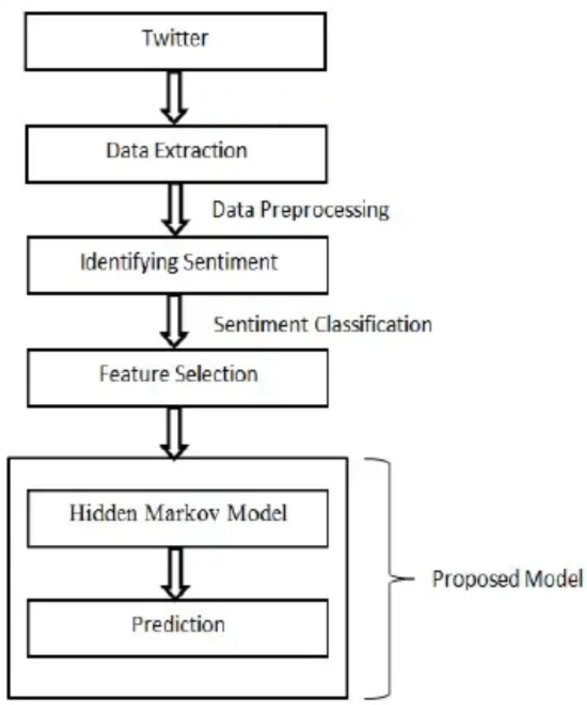

Figure 2: Test Architecture

Our project aim to solve the following questions: 1. Does Early detection possible in the case of doge coin? 2. Does sudden doge coin rise/fall during 2021 could've predicted using HMM output? 3. Does Covid and Influenza have any effect on doge coin price surge?

4.2 Experiment 1 - Sentiment Analysis On Twitter Data

In order to get the sentiment scores from the twitter data, we first collected 100000 tweets based on hashtags related to dogecoin for the days of March 15 - Sep 15. Being the raw data, this required further processing, since there is noise in the tweets that needs to be filtered out. The tweets have combinations of expressions, URLs, symbols, emoticons, and user's mentions. Because of such causal nature of twitter usage by people, it gives rise to a noisy dataset which requires a structuring for it to be learned by a classifier. There are a certain number of general pre-processing steps performed for this in order to apply sentiment analysis on the tweets and quantify them with a score that can be used to correlate them with the Dogecoin prices. The steps involve -

- Removal of duplicate tweets and converting the text to lowercase
- Tokenizing the data and removing stopwords
- Lemmatizing and stemming
- Removing the symbols and special characters

For this, the readily available pre-processing packages were used as well as regular expressions, to remove the tags , quotes and question marks as they cause biased results for sentiment analysis. We used the combined lexicon and rule-based sentiment analytic software Valence Aware Dictionary and Sentiment Reasoner (VADER) [Hutto (2014)]. It is capable of both, detecting the sentiment intensity in text and the polarity(positive, negative, neutral). Tweets with a positive opinion are assigned Positive values while negative values are assigned to tweets with a negative opinion. Being tailored explicitly to the sentiment of social media, VADER is generally expected to show positive results. The polarity score is between -1 and 1, where -1 to 0 is negative, 0 to 1 is positive and 0 is neutral.

Observations and Analysis

After assigning a score via sentiment analysis to every tweet, two time series indicating the opinions over Twitter related to DogeCoin were created - one each for the positive and negative opinions. Since the two different types of sentiments may have an asymmetric effect on the prices, they were plotted in the below figure along with the observed DogeCoin closing prices for the same time period, beginning March 15 (indexed 0).

As you can see from the image there is a sudden increase in the doge coin price around May 15(indexed around 50), this is very unexpxected behaviour. We attempted to study this behaviour using only sentiment analysis of twitter data first. But, as you ca see from the figure 3, we can't be able to find any correlations in the data. This explains the pandemic like behaviour of doge coin market. However, there are minute correlations in the sentiment data, we can estimate those correlations, only when combined with the early detection approaches from Epidemiology. Hence, we wanted to use social media data along with the HMM model to detect hidden states. In our next experiment we used an HMM model to overcome the pitfalls of this experiment.

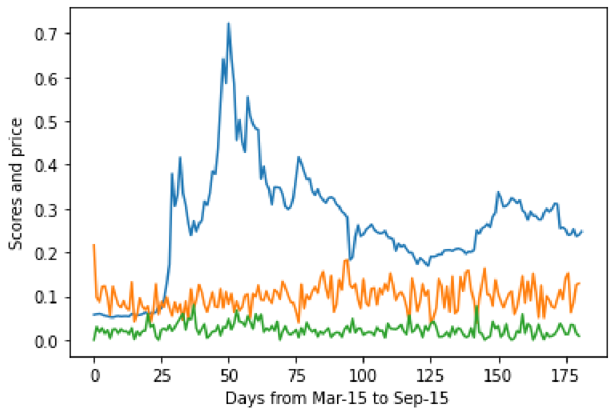

Figure 3: Twitter Sentiment and Doge Coin Prices

4.3 Experiment 2 - HMM on Social media Analysis

In this experiment, we wish to achieve better prediction results using an HMM model. First, we used twitter positive sentiment to train the HMM model with five hidden states with each state representing each stage in an crypto bubble. Since, HMM follows expectation - maximization, we train the model for different iterations to prevent it from falling into local maxima. After that, we trained another HMM model using twitter negative sentiment analysis, we repeated the same steps as of positive sentiment HMM model to prevent it from falling into local maxima. After that we ran predictions on both the models, the following results are plotted in figure 4.

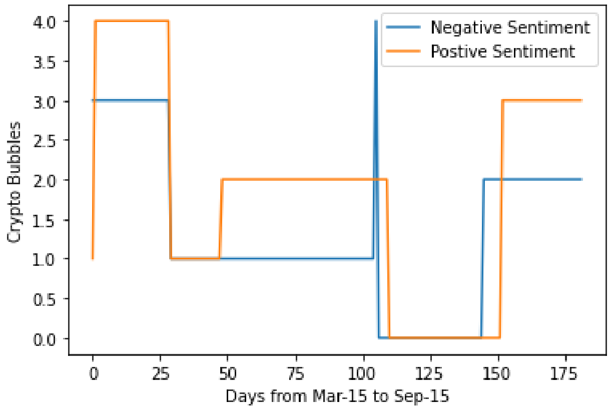

Figure 4: Twitter Positive and Negative Sentiment trained on HMM

Observations and Analysis

The output consists of time series data on the x-axis and it's corresponding hidden states on the y-axis. We plotted both the graphs for negative sentiment and positive sentiment on the same graph to show the correlations. We can infer the state a particular date is in from the graph, for example let's consider the date when doge coin had a sudden rise, it can be seen from the graph 3, the doge coin's boom happened around index 50 (May 10). Around this time, from figure 4 we can observe that the graph for positive sentiment entered the hidden state 2, while the graph for negative sentiment entered the hidden state as well. Such early detection of doge coin's rise is useful in detecting the booms of the market, which couldn't have been done with the normal regression models. This observation answers the main question of Does Early detection possible in the case of doge coin?. After that if we observe the graph again, around index 100 there has been a sudden fall in the twitter sentiment values, i.e., the dates from 105 to 150 was in stage 0. There is also a sudden fall of negative sentiment from state 4 to state 0, and positive sentiment fell from State 2 to State 0, if we again observe the doge coin price chart from figure 3, around index 100, the price dropped. So, using HMM models, this could've predicted earlier as well, and another example is from the figure 4, at 150 index both the graphs for positive sentiment and negative sentiment entered the hidden states of 2 and 3 again, and as expected the price of doge coin went up. This behaviour answers our second question Does sudden doge coin rise/fall during 2021 could've predicted using HMM output?. We put more focus on 2021 because doge coins' market expressed more pandemic like behaviour in 2021. Hence, based on the above analysis we can say doge coins sudden rise/fall in prices during 2021 period could be predicted to certain level. This is only possible because of the earlier works in epidemiology, it provided the idea of early detection using the hidden states technique in HMM.

4.4 Experiment 3 - HMM on Covid US data and Influenza data

Additionally, we wished to study the influence of covid US and influenza data on doge coins' market. We wish to achieve a practical framework using HMM that may be used in the future for analyzing external factors' influence on doge coin market. For that purpose we trained two separate HMM models again for both Covid and Influenza data. First a new HMM model is created with hidden states, then covid new infections data from the same period of 6 months is trained on it. Similarly influenza data for six months is collected and trained using HMM to detect hidden stages in the infections. The results plotted in figure 5 below.

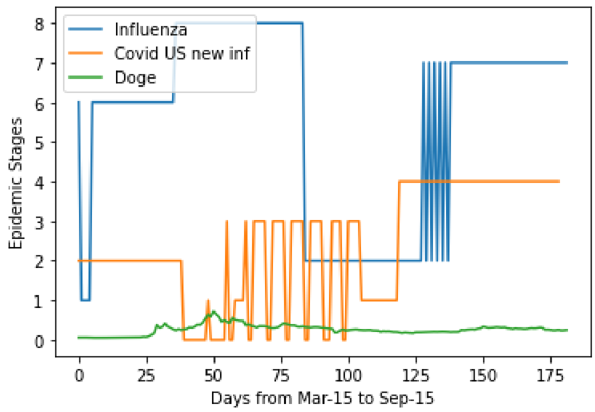

Figure 5: US covid inf data and Influenza data trained on HMM

We can certainly say that influenza data has no effect on doge coin prices, but in first glance it appears that covid data has some correlation with the doge coin prices, after careful observation, there are lot of fluctuations in the states of covid data new infections, which means an HMM model couldn't be able to detect the hidden state of covid pandemic, so it's hard to predict the rise of new covid cases early. Hence, we can say covid and influenza data has no effect on the doge coin prices.

5 Conclusion and Future Scope

The number of cryptocurrencies has increased manifolds over the past decade. It can be very hard to pick one specific currency to use, and that can be attributed to the vast diversity of coins. There are always concerns around the dynamic pricing of these currencies as well as the popularity. Herding behaviour is observed to be very common in Cryptocurrency prices and people often share their opinions to strongly perturb the market. Consequentially, as a part of the popular culture, public figures associate themselves with the Cryptocurrencies and are often termed Tastemakers as they endorse them.

This work highlights how the techniques of epidemic detection can be applied to Tweets for prediction of the Cryptocurrency price bubbles and there is also a significant observation that like the spread of an infectious disease , ideas of investments also spread and form bubbles. In order to exhibit this, we performed the training of an HMM model based on the sentiment scores calculated for the cleaned tweets extracted using TWINT library. Upon observation, the epidemic and non-epidemic stages were learned and classified by the model, however when trained using the influenza and COVID infections data, the model didn't perform that well. This helped us conclude the strong correlation between the DogeCoin prices and the behaviour expressed by users on the internet, often via Tweets.

We hope to create trading strategy incorporating the output generated from Hidden Markov Model and predict the next price boom or burst in the Dogecoin prices. This would help in determining and proving the utility of state probabilities by converting it into a financially beneficial trading technique which would perform much better than a similar benchmark.

References

H. D. Dwivedi A. Jain, S. Tripathi and P. Saxenae. 2018. Forecasting price of cryptocurrencies using tweets sentiment analysis. *Eleventh International Conference on Contemporary Computing (IC3)*.

T. Shtatland E. Shtatland. 2008. Another look at low-order autoregressive models in early detection of epidemic outbreaks and explosive behaviors in economic and financial time series. *NESUG'20 Proceedings, Paper 363*.

E.E. Hutto, C.J. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.*

C. Bilanakos I. Georgoula, D. Pournarakis. 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *SSRN Electronic Journal*.

Xiaojun Zeng Johan Bollen, Huina Mao. 2011. Twitter mood predicts the stock market.

Wolfgang Kuhle. 2018. Thought viruses and asset prices.

Carrat F. Le Strat Y. 1999. Monitoring epidemiologic surveillance data using hidden markov models. *Statistics in Medicine, 18(24):3463–3478*.

A. Lopez-Quilez M. A. Martinez-Beneito, D. Conesa and A. Lopez-Maside. 2008. Bayesian markov switching models for the early detection of influenza epidemics.

Ross C. Phillips and Denise Gorse. 2017. Predicting cryptocurrency price bubbles using social media data and epidemic modelling.

Xindong Wu Saeed Abdullah. 2011. An epidemic model for news spreading on twitter.