Prediction of Stock Exchange Price Bubbles using Epidemic Modelling on Social Media and News

Algın Poyraz Arslan

School of Computer Science

College of Science

Izmir Institute of Technology

Submitted in partial satisfaction of the requirements for the
Degree of Bachelor of Science
in Computer Science

Supervisor Assistant Professor Selma Tekir

May, 2021

We approve the thesis of Algın Poyraz Arslan

27/05/2021

.....

Assistant Professor Selma Tekir

Supervisor

Department of Computer Engineering

Chapter 1

Introduction

This thesis focuses on mathematical modelling of the speculation of financial instruments, specifically stock prices, that occurs on social media. Considering how news or a popular topic spreads over a network like a disease, an epidemic model has been considered in this thesis. The main objective is to display the correlation or causation between epidemic modelling which uses compartmental model SEI on social media and stock price. This causation may help one to predict if a stock price is a bubble or not. Epidemic models using Hidden Markov Models (Phillips and Gorse, 2017) have been proved to be useful for predicting financial bubbles with cryptocurrencies.

A compartmental model is simply a mathematical demonstration of a population where an infectious disease occurs. The population is separated into compartments such as Susceptible, Exposed, Infected, Recovered, etc. Population in each compartment changes with different rates at a unit of time, which allows us to solve the differential equation system with these rates and simulate the population's state in a distinct time.

The thesis also includes a selection process among different compartmental models to yield a better model. Considering different approaches and research results, the SEI model results are discussed within the thesis. Future work to the thesis would be building a trading agent using this model and comparing the profits of the agent to a base and profits of Phillips and Gorse, 2017.

Chapter 2

Literature Search Results and Related Works

2.1 Model

To the best of our knowledge, the first ones to use epidemic modelling for detecting financial bubbles are Phillips and Gorse, 2017. Although it is proposed that a compartmental model such as SIR could be used, the paper mostly discusses financial bubbles on cryptocurrencies based on the Hidden Markov Models. As complementary work in this thesis, we have researched based on the compartmental models. As future work, since the work is complementary, using their Phillips and Gorse's paper's result to compare this research's results should be done.

Using a compartmental model comes with its problems. Defining a population and compartments, differentiating infection rates and their usage, etc. Stai et al. (2018) introduces the concept of the temporal dynamic which is basically a popularity analysis over the network of news over time. They claim that constant infection rate usage for simulation of a model is solid if the news popularity fluctuates over time but using time varied infection rates decreases the error for non-fluctuating news. Our project also includes a popularity analysis over a period of time to ensure that constant infection rates are the solid case.

Even though an SEI model has been chosen for the project a recovered compartment has been included. According to the Strahilevitz, Odean and Barber investors repeat actions that previously resulted in pleasure while avoiding actions that previously

led to pain. Future work may be compared to this thesis where a recovered state is included in the population for investors that have suffered a loss in the stock. Unfortunately, this model and methodology are out of the scope of our project.

Kumar et al. (2020) use the SEI model in their paper which is used in our project as well. The paper treats whole network as susceptible, anyone that has posted, replied, retweeted, mentioned the related topic as infected, anyone that has a connection with an infected node as exposed to define the population. Differently from their paper, we have included favorite as an infection spread method on an infected post. Paper also includes a self-infection rate which is logical for financial news because they can be encountered outside of the network easily.

2.2 Social Media

The paper mentioned above (Phillips and Gorse) uses solely reddit[1] as medium to process social media inputs. This may be a good approach for cryptocurrencies which are separated in Reddit with different subreddits for each cryptocurrency they have inspected. It is not the case for stocks. Each stock does not have a separate subreddit for them. This makes it hard to use inputs used in the paper of Phillip's and Gorse's. A good choice may be also twitter[2] where more data can be acquired. Related research is done on Twitter by (Vosoughi, Roy and Aral, 2018). The paper focuses on how to determine the state of users in social media according to their behavior. Even though paper took retweets as the only option to the spread of rumors, considering twitter medium, further action may be taken to engage with news and spread it.

Another paper that inspects the news spreading over twitter medium is provided by Abdullah and Wu (2011). They broadly discuss models such as SI, SIR, SIRS, SEIR and concludes that SIR is the best model for Twitter. Even though they conclude on SIR we have chosen SEI for the project considering the news is niche and may require time to have research out of network for a node to get infected. This latent

[1] www.reddit.com
[2] www.twitter.com

time can be modeled with an exposed state. The work treats each node as infected when they publish a tweet and treats each follower as susceptible. The article also mentions a basic reproduction rate (R0), if, according to the authors, it is below 1 epidemic will not be turning into a pandemic. This rate may be the indicator for the trading agent for follow-up work in this project.

Another paper that inspects compartmental model usage on Twitter and uses an exposed state within is provided by Jin et al. (2013). The article Jin et al. considers every user susceptible and considers infected if a person has tweeted about the subject. Every post that has been posted by the infected, exposes some new nodes in the network. That seems logical with an exposed state for this project. A susceptible user may be exposed to stock and research the stock then decides to invest in the stock. The authors also introduce a new compartment which is skeptics. The people that never tweet about the stock. Skeptic compartment is solely encountered in this article and is thought to complicate the model unnecessarily, and omitted from this project. Future work may include a skeptics compartment to analyze the results and compare them with the current one.

Chapter 3

Method

3.1 Data Colletion

The problem at hand required a nontrivial data collection effort. The data for the Twitter network is collected from github[1] and selected due to the long time interval and relatively high discussion rate on the network. Data is provided by Xu and Cohen (2018).

The simulation is required to initialize the network population and average follower variables. This data is acquired from statista[2] for the years between 2013-2016, and an average of them is calculated.

To make a relative comparison between a bubble and non-bubble stock, one-month data of WallStreetBets from Reddit is collected where posts have caused a price bubble over Gamestop stocks within a month (Royt, 2021). Data is acquired from kaggle[3] which is posted by Gabriel Preda.

The data of daily average comment per post for Reddit and the data of average subscriber of WallStreetBets is acquired from subredditstats[4]. The average daily active user in WallStreetBets is calculated by. $(Daily active user of Reddit/Total user of Reddit * Subscriber of WSB $). Network data of Reddit is acquired from backlinko[5].

[1]https://github.com/yumoxu/stocknet-dataset
[2]https://www.statista.com/statistics/303681/twitter-users-worldwide/
[3]https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts
[4]https://subredditstats.com/r/wallstreetbets
[5]https://backlinko.com/reddit-users

h

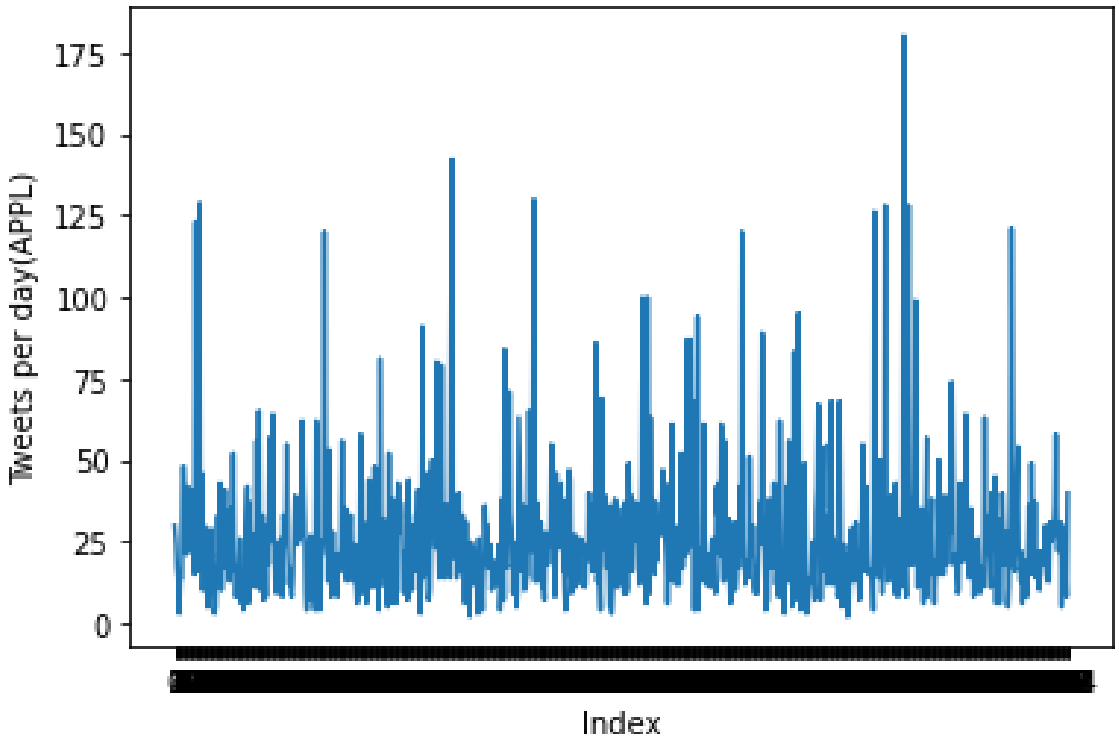

Figure 3.1: Popularity of Apple stock over time

The data for fundamental comparison is obtained from investing[6]

3.2 Data Analysis

From the collected Twitter data, sample stock data is chosen for the sample experiment (AAPL). The stock is chosen to have high volume and high recognition to increase the chance that stock prices are never bubbled. Chosen data has stripped of irrelevant information, remaining simply the date of tweets and favorite counts. No outlier is checked and removed from the data considering volumes of trading highly fluctuates depending upon the economical conditions over the globe. Data is also analyzed for popularity over time in the network so that it is correct to use constant infection rates. Analysis showed that Apple stock's popularity is fluctuating over time which means there is no need for time varied infection rate usage.

The Reddit data has stripped of posts of stocks other than Gamestop. Each post's

[6]www.investing.com

title and body is examined and selected with the Levenshtein distance of 1 with the keywords as (gamestop,game,gme,$gme). Selected post scores are considered as favorites in the Twitter network and used as an infection spread method as well.

The Levenshtein distance algorithm compares two strings and counts the number of inserted, deleted, or substituted characters. The counted operations are considered as Levenshtein distance. This distance helps us to typos, misspellings in the data.

3.3 Experimental Method

Previously mentioned twitter data of Apple stock and Reddit data of Gamestop is fed into an SEI model as ordinary differential equations. The problem of obtaining exposure, self-exposure, tweet or post rate of already infected node, and average out-degree of a node has been solved with initial values. The sum of squared error is calculated from the difference of infections and it is minimized to find out real-world parameters. Ensuring the avoidance of local minima considered with the usage of basin-hopping. We have also used a brute force analysis on a range for the average out-degree of a node to ensure the best results. The population of Twitter and Reddit was assumed to be not changing and calculated as an average in the time interval of the data.

The system is solved once again with real-world parameters to identify infected, exposed and susceptible at a certain time. For future work, we plan to develop a trading bot regarding the reproduction rate(R0) analysis on the data set. These R0 rates over time would be analyzed with the stock's price in the same time interval.

Results of the simulations can be seen in 3.2 and 3.3. In the Reddit network, about 1/4 of people are infected with Gamestop stock and about 1/3 is exposed to it. On the contrary in the Twitter network, apple stock news are quite non-contagious considering only about 0.00007% of the network is infected.

Fundamental analysis of Gamestop does also show that stock ratios are quite high from similar stocks in the same sector that is chosen in different locations over the

h

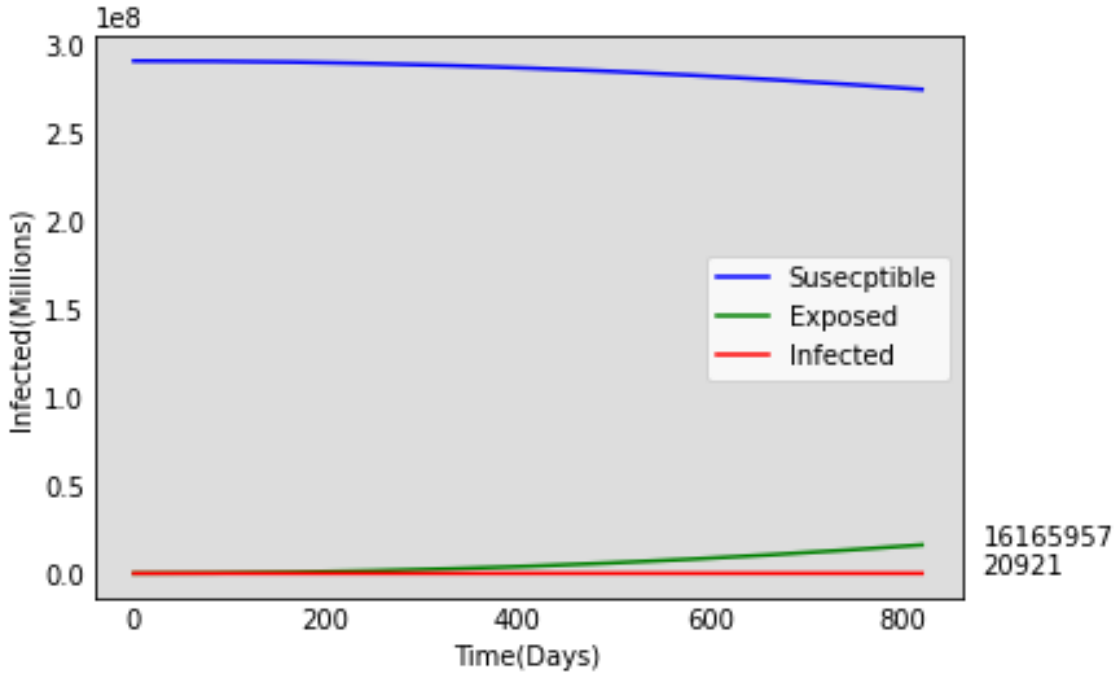

Figure 3.2: Population change of Apple stock in network

h

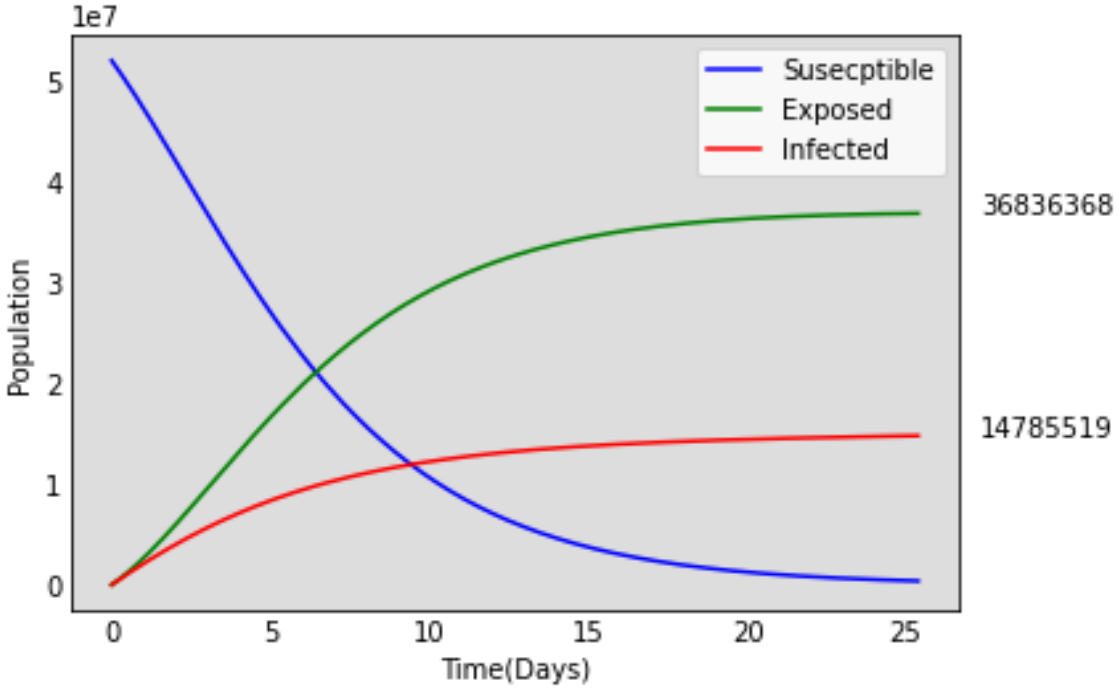

Figure 3.3: Population change of Gamestop stock in network

globe. Even after the burst of the bubble, one month later ratios lowered significantly but they were still higher than the average. This can be seen in tables 3.1 and 3.2

Table 3.1: 27 January 2021

Name	Code	Price/Earnings	Price/Book	Price/Sales	Price
Gamestop	GME	NAN	68.21	4.67	347.51
Walmart	WMT	30.03	4.95	0.72	143.84
Amazon	AMZN	77.42	17.41	4.21	3232.58
Target	TGT	23.62	6.70	1.01	178.28
Best Buy	BBY	16.60	6.41	0.62	113.57
Ceconomy	CECG	4.574	0.88	0.04	5.34
Frasers	FRAS	19.30	1.71	0.59	433.8
Fnac Darty	FNAC	NAN	0.87	NAN	45.14
Average	Average	28.59	13.39	1.70	

Table 3.2: 26 Februray 2021

Name	Code	Price/Earnings	Price/Book	Price/Sales	Price
Gamestop	GME	NAN	19.97	1.37	101.74
Walmart	WMT	27.43	4.52	0.66	131.37
Amazon	AMZN	74.08	16.66	4.03	3092.93
Target	TGT	24.31	6.9	1.04	183.44
Best Buy	BBY	14.67	5.66	0.55	100.35
Ceconomy	CECG	4.42	0.85	0.04	5.17
Frasers	FRAS	20.88	1.85	0.64	469.2
Fnac Darty	FNAC	NAN	0.97	NAN	50.25
Average	Average	27.63	7.17	1.19	

System is solved by odeint function from scipy library[7]. Plottings are done by matplot library[8]. Data is processed with the help of pandas[9] and numpy[10] library.

[7] ww.scipy.org
[8] www.matplotlib.org
[9] www.pandas.pydata.org
[10] www.numpy.org

Chapter 4

Experimental Setup

4.1 Libraries

Used libraries for the article are:

- scipy
- numpy
- matplotlib
- pandas
- python-Levensthein

These libraries can be installed with the help of the pip module with the following command: pip install scipy numpy matplotlib pandas python-Levenshtein.

4.2 Initial Parameters

Initial parameters for the Reddit network can be found below:

- Average out-degree: 6
- Susceptible :52000000
- Exposed:0
- Infected:0
- Alpha:0.5

- Beta:0.8
- Gamma:0.1

Average infection per post in interval of (28.01.2021-22.02.2021) respectfully are [5.77, 22.22, 73.43, 78.49, 34.5, 13.91, 24.36, 23.76, 18.20, 28, 30.81, 98.24, 222, 208.33, 197.54, 190.9, 91.6, 69.11, 120.98, 20,50.73, 26.5, 31.96, 27.15, 34.13, 38.78]

Initial parameters for the Twitter network can be found below:

- Average-out degree:300
- Susceptible:291000000
- Exposed:0
- Infected:0
- Alpha:0.1
- Beta:0.2
- Gamma:0.8

Chapter 5

Conclusion-Impact & Future Work

The article shows that a bubble stock can be identified from discussions over social networks. Also, the epidemic model method differentiates contagious and non-contagious stocks which may help one in decision making for investment opportunities.

Future work for the project includes comparison with different compartmental models with recovery compartment such as SEIR, developing a trading bot, and discussing the profit made with decision making using the article's model.

References

Abdullah, Saeed and Xindong Wu (2011). 'An Epidemic Model for News Spreading on Twitter'. In: *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pp. 163–169. DOI: 10.1109/ICTAI.2011.33 (cit. on p. 3).

Jin, Fang et al. (2013). 'Epidemiological modeling of news and rumors on Twitter'. In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis - SNAKDD '13*. ACM Press. DOI: 10.1145/2501025.2501027. URL: https://doi.org/10.1145/2501025.2501027 (cit. on p. 4).

Kumar, Sanjay et al. (2020). 'Modeling Information Diffusion In Online Social Networks Using SEI Epidemic Model'. In: *Procedia Computer Science* 171, pp. 672–678. DOI: 10.1016/j.procs.2020.04.073. URL: https://doi.org/10.1016/j.procs.2020.04.073 (cit. on p. 3).

Phillips, Ross C. and Denise Gorse (2017). 'Predicting cryptocurrency price bubbles using social media data and epidemic modelling'. In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7. DOI: 10.1109/SSCI.2017.8280809 (cit. on pp. 1–3).

Royt, Alex (Jan. 2021). *The Gamestop bubble is an age-old financial craze with a modern twist*. [Online; posted 29-January-2021]. URL: %5Curl%7Bhttps://www.washingtonpost.com/outlook/2021/01/29/gamestop-bubble-is-an-age-old-financial-craze-with-modern-twist/%7D (cit. on p. 5).

Stai, Eleni et al. (2018). 'Temporal Dynamics of Information Diffusion in Twitter: Modeling and Experimentation'. In: *IEEE Transactions on Computational Social Systems* 5.1, pp. 256–264. DOI: 10.1109/TCSS.2017.2784184 (cit. on p. 2).

Strahilevitz, Michal Ann, Terrance Odean and Brad M. Barber (2011). 'Once Burned, Twice Shy: How Naive Learning, Counterfactuals, and Regret Affect the Repurchase of Stocks Previously Sold'. In: *Journal of Marketing Research* 48.SPL, S102–S120. DOI: 10.1509/jmkr.48.SPL.S102. eprint: https://doi.org/10.1509/jmkr.48.SPL.S102. URL: https://doi.org/10.1509/jmkr.48.SPL.S102 (cit. on p. 2).

Vosoughi, Soroush, Deb Roy and Sinan Aral (Mar. 2018). 'The spread of true and false news online'. In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559. URL: https://doi.org/10.1126/science.aap9559 (cit. on p. 3).

Xu, Yumo and Shay B. Cohen (July 2018). 'Stock Movement Prediction from Tweets and Historical Prices'. In: *Proceedings of the 56th Annual Meeting of the Associ-*

ation for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, pp. 1970–1979. DOI: 10.18653/v1/P18-1183. URL: https://www.aclweb.org/anthology/P18-1183 (cit. on p. 5).