Social media Effects on the market: Reddit Data analysis on Stocks

Preprint · May 2021

DOI: 10.13140/RG.2.2.24180.88960

1 author:

Juan Andrés Talamás
Tecnológico de Monterrey
1 PUBLICATION **0** CITATIONS

SEE PROFILE

Social media Effects on the market: Reddit Data analysis on Stocks

Juan Andrés Talamás Carvajal
Tecnologico de Monterrey,
School of Engineering and Science,
Monterrey, N.L.
juan.talamas@tec.mx

***Abstract*—As social media continues to become a more essential part of life in general, including hobbies and general interests, it is beginning to have effects that were previously not considered possible.**

While the social and sale effects of online communities is well known, these are usually focused on individual choice. (i.e. what to buy, where to go eating, etc.) Actually affecting the value of a company was something that almost no-one expected. The recent events regarding Gamestop, Dogecoin, Blackberry, and some others, proves otherwise and shows that social media has reached a point where these communities can consciously and purposely increase or decrease value.

This analysis will attempt to explain the overall influence of social media over perceived value (stock price), and attempt to find a model that could lead to better understanding of digital communities and their effects on the stock market.

***Index Terms*—Social media, Twitter, Reddit, stocks, Data science**

I. Introduction

This article will follow the CRISP-DM methodology as explained in [1]. As such, it includes sub-sections that correspond to the different components of the model cycle. CRISP-DM stands for Cross-industry standard process for data mining, and is an open standard process model that describes common approaches used by data mining experts.

One of the most fundamental characteristics of the data-driven society that we now live on is social media. It has grown to the point that it is possible to find groups and conversations about any and all aspects of daily life (Cooking, school, hobbies, general interests, ect.), and more recently, even highly specialized forums for more specialized themes [2].

What began as a way to share what people thought to small groups consisting usually of real-life friends and family has now evolved to a platform where something as distant from daily life as the stock market is now a topic that is easy to find, and provides ample information regarding that topic [2].

More recently, it appears that digital life has started to bleed into real life, as a series of groups dealing with the stock market seemed to successfully bring up the value of several stocks that, according to the general ideas of how value is determined, should not be nowhere near as high as they climbed [3].

The stock market has been a subject of study for some time now, with an emphasis on predicting the prices as they rise or fall as closely as possible. Stock market profits depend almost exclusively on the future performance of the markets, which has generated great interest by the machine learning and data mining community. Classic approaches have so far proven to be insufficient, as the markets have proven to be much more complex than was expected. Some recent approaches have attempted to use hybrid models, either by combining 2 existing prediction algorithms in sequence [4], merging a Convolutional Neural Network or stock prices with the results of a Sentiment Analysis technique [5] [6]. However, most of the existing literature on the topic uses a single Machine Learning algorithm, but seeking to optimize the different parameters used. The most common are Support Vector Machines [7] [8] [9], Convolutional Neural Networks [9] [10], or autoregressive methods as are AR, ARMA, and ARIMA [11] [12].

However, this subject has so far escaped a true solution as stock markets have an inherent random nature to them, which stems from human perception of the company, business cycles, media coverage, or even social media involvement. The research presented in this article will show that, as we become more interconnected and social media continues to grow, the conversations and group decisions made online have a continually increasing potential to affect real life markets. A few years ago, no-one would have believed a relatively small group of people treating Wall Street as a casino would have any weight on a company's value, but as the Gamestop story shows, it became a reality.

II. Methods and Data

The method followed for this research is the Cross-industry standard process for data mining (CRISP-DM), which is a staple in the data mining community and commonplace when applying data science to business applications. CRISP-DM is compromised of six major phases, which are: Business Understanding, Data Understanding Data Preparation, Modeling Evaluation, and Deployment. While the name of the phase is "business" understanding, this refers to the comprehension of the target problem, and along with data understanding, it corresponds to the initial phase of the project. After the data has been prepared, we advanced to the data preparation and modeling cycle, in which we prepare our data for the model, fit it and evaluate our results. After evaluation, there is only the deployment phase, where an evaluated model would be

used outside of a ”lab” environment to make real decisions and predictions.

In this particular project, four different data sets were obtained: A kaggle data set that contains Reddit posts from the subreddit WallStreetBets (a forum for stock enthusiasts), which has been credited with being the group mainly responsible for the abnormal price variations researched in this article, and 3 Nasdaq data sets containing historical values for the high, low, starting, and ending prices for AMC (AMC Theatres, a cinema company theater chain), GME (Gamestop, a video game retail chain), and NOK (Nokia, a telecommunications company).

The first phases of CRISP-DM will be explained here, while the modeling and evaluation will be covered in the results section of the article.

Business understanding: These 3 companies were performing poorly and were targeted for ”shorts” in which investors bet that the stock will continue to fall. The Wallstreet Bets (WSB) subreddit found out and attempted to perform a ”short squeeze” in which they purposely build up the price of a stock because ”shorts” have an expiration date at which the stocks have to be bought regardless of the price, making it very profitable for the subreddit if they succeed.

Data understanding: The three Nasdaq data sets are .csv files that contain the date, Close/Last price, Volume of the stocks traded, Opening price, Highest price, and Lowest price for their corresponding stock for 6 working months (holidays and weekends are excluded due to the stock market being closed). All values except date are numerical. The WSB data set contains the title, id, url, number of comments, number of the particular post, body, and timestamp of the comment.

Data preparation: In order to obtain meaningful data from the WSB data set, we first need to comb through the comments to find relevant features. In this case, we look for instances of a selected stock per day, either in the title or body of the comments, the amount of comments relating to the different stocks, and the amount of up votes (score) of relevant threads. A high amount of comments or up votes indicates larger user interaction regarding the stock for all features mentioned. As Reddit does not show publicly the amount of views for any given comment, these values correspond were used instead. This was first performed by comment and then aggregated per day. The values per date were then merged with the stock data sets to create 3 different data sets, each corresponding to a stock. Finally, these 3 were then concatenated to form the panel data that was used. The final set was arranged by date and stock name, and contained the following variables: Body count, Title count, number of comments, total score (up votes), Closing price, Volume traded, High price , and Low Price. Before going into the modelling phase, we performed informative attribute selection on our data. The heat map showed below in “Fig. 4” resulted in dropping the ”title” column as it is highly correlated with the ”body” values, but contributed less on our target variables. Further selection was performed by way of backwards elimination, in which every other attribute except ”comm” (Number of comments) was discarded.

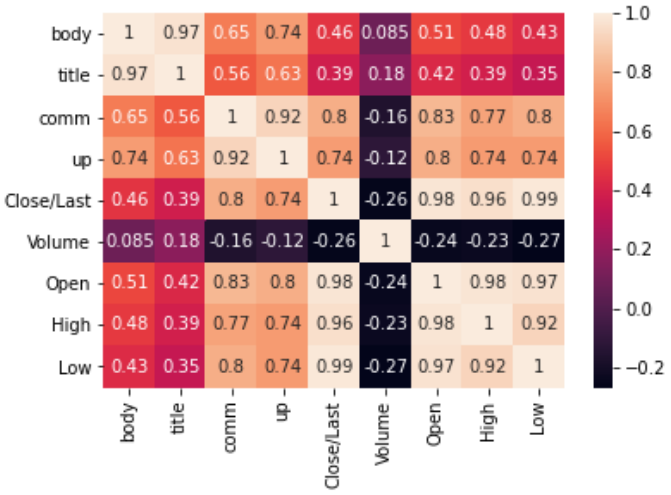

Fig. 1. Heat map indicating correlation between attributes.

III. Results

Modeling: As the data has a date associated with it, the initial models tested belong to the Panel Data family. The initial test with Pooled OLS is generally there just to serve as a baseline for further testing, and is used to validate the use of a panel data strategy. However, we found out that the tests instead pointed to the fact that Pooled OLS was in fact the better model to use. The test used to test the assumptions for Pooled OLS (linearity, mean of residuals equal to 0, multicollinearity, homoskedasticity, etc.) showed p-values larger than 0.05 (White test, Breusch-Pagan Test, Durbin-Watson Test), and as such, a different approach than panel data was needed.

While Panel data is not the ideal family of models to use, the fact remains that the original data set has a time component. Due to this we can perform a time series analysis instead of panel data one as a way of incorporating the time aspect of the data to our model.

What we encountered when doing the time series analysis was the following: It appears that the time series for the different stocks are not stationary, and as such, we require to difference the data first. “Fig. 4” shows the GME data after being decomposed into its components. We can observe that the trend follows almost exactly the overall values, and that seasonality is small compared to the true values.

For the GME case, an ARIMA model with p=0, d=1, and a=1 was performed after the values were obtained from the auto.arima function in R. While the model passes the auto-correlation and partial auto-correlation tests after a few points, and the box test had a p-value much bigger than 0.05, the forecast errors do not follow a strict normal distribution, as can be seen on “Fig. 4”. According to the empirical rule of normal distribution (3 sigma or 68-95-99.7 rule), our errors should be more distributed under the curve, but it appears that

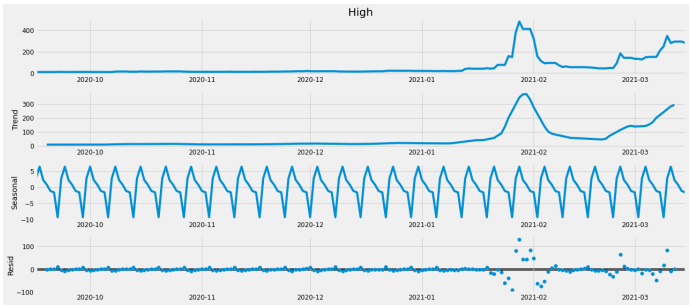

Fig. 2. Decomposition of GME "high" stock values.

we have a heavy disposition towards the zero or close to zero errors.

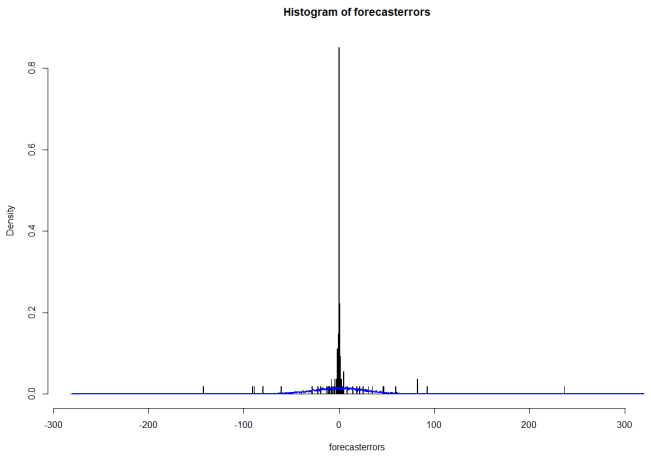

Fig. 3. Histogram of forecast errors for GME ARIMA(0,1,1) model.

As the data panel approach was not appropriate, we separated the Panel data into 3 smaller sets, each one including only the data for one stock.

The forecast for the GME "high" prices from the ARIMA(0,1,1) model can be seen below in "Fig. 4". Some important things to observe are that the model would be incapable of predicting the big "jumps" even when relaxing the confidence of our predictions. While disappointing, this is not unexpected, as those spikes correspond to the dates where most of the planned activity from WSB was taking place. As these data points do not correspond to a time-based phenomenon, but a human interaction one, a time series analysis may not be the best tool in this case.

Following the time series results, we then attempted regression algorithms for each of the different stocks.

Initial testing with regression algorithms started with OLS, as it still serves as a good baseline when working with linear regression models. In order to Find the best model, we compared the adjusted R-squared values for a linear model, a polynomial model, and a support vector regression model for the 3 different stocks. Each case used training and testing sets to ensure the validity of each model. "Table III" below shows the results for all cases.

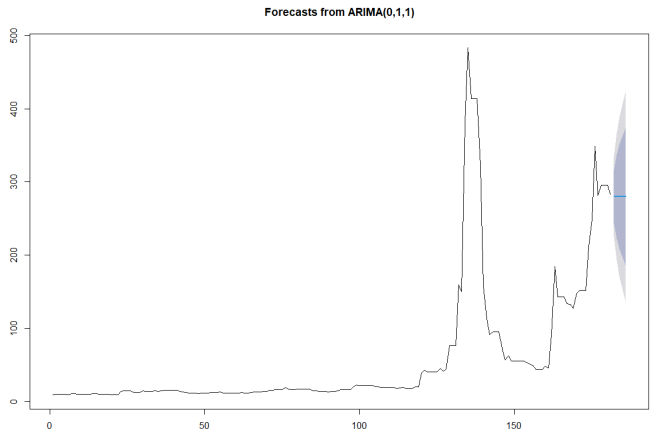

Fig. 4. Predictions 5 next values for GME ARIMA(0,1,1) model.

TABLE I
SCORES OF LINEAR MODELS FOR EACH STOCK

Stock Name	Linear Regressions		
	OLS	*Polynomial LS*	*SVR*
GME	0.417	0.099	-0.041
AMC	0.245	-0.123	-0.027
NOK	0.672	0.488	0.478

From the table, we can observe that OLS continues to be the best model for all the stocks included on the analysis. In order to have a better understanding of the models, and to ensure the quality of our predictions, we used cross-validation techniques to obtain the average test and train scores of our 3 stocks. The results can be seen in "Table III" and "Table III".

TABLE II
TRAINING SCORES FOR CROSS VALIDATION RUNS

Stock Name	Cross validation run		
	1	*2*	*3*
GME	0.387	0.212	0.315
AMC	0.069	0.227	0.552
NOK	0.013	0.528	0.528

TABLE III
TESTING SCORES FOR CROSS VALIDATION RUNS

Stock Name	Cross validation run		
	1	*2*	*3*
GME	0.198	-2.542	-1.022
AMC	-3.133	-0.955	-4.120
NOK	-0.570	-0.975	-1.223

IV. DISCUSSION

Evaluation: After running the cross-validation procedure, we could observe that the scores between the test and training scores are drastically different. We usually expect the test sets to obtain slightly worse scores due to the very nature

of the test-train split, but the results presented before show values that go into the negatives for the OLS models we ran. The score used in this analysis (R-Squared) returned negative values, meaning that the model would return worse results than running a straight line through the values.

Throughout the study, the data has shown to not have the necessary characteristics for panel data analysis, time series analysis, and most of the initial informative attributes that we planned to use for and ordinary least-squares regression were deemed not optimal after a backwards elimination procedure. While there appears to be a correlation between the amount of traffic and comments in the subreddit "Wall Street Bets" and the sale prices of stocks, it appears that the relationship is not one that can be used to properly model the behaviour of stock prices to an acceptable degree.

The high variance between runs could be an indicator of data points that do not follow the general trend of our data-set, and the inclusion or exclusion of those points from the training set heavily influences the results. It is likely that in this case, the target data points (i.e. the highest values) are outliers even when taking into account the increased traffic seen on the forums.

One possible reason for these results is the social nature of our attributes. While there was a measurable and easily proven increase in traffic, comments, and up-votes on the subreddit, participation in the reddit community does not ensure participation on the "short" of these stocks. People might have gone to the forums just to see what was happening, and not committed to either buying or holding the stock. Another case would be that of the opposite: people who heavily participated but did not comment at all on the forums. Social media usually follows a pattern where comments or up votes are much less represented than views. It could be the case that a number of people followed the discussion online and participated by buying and selling, but did not interact with other participants.

V. Conclusion and future work

We attempted to generate a model that would predict stock prices for the GME, NOK, and AMC companies using several different features from the social media site Reddit. As the data was generated daily, we attempted both a data panel and time series approach, but the models generated did not comply with the necessary characteristics for either of those startegies. An ordinary least squares regression (OLS) was then attempted, but the obtained models, while complying with the necessary assumptions, did not provide satisfactory scores when using the testing set (R2 values going into the negatives). Due to this, we would not go on to deploy the model, but instead cycle back in the methodology and continue working.

A possible explanation of this phenomenon is the nature of our data: Social media comments or up-votes do not usually show the true number of people viewing the content, and it is possible that the people that did participate in the forums did not do so when buying stocks, or vice-versa.

Stocks have proven to be one of the most difficult to predict values in modern history, mostly due to the fact that the perception buyers or sellers have of the company is something that is not trivial to evaluate. A number of factors that are not dependant on the performance of the company also influence prices, and in some cases, like the one mentioned in this article, heavily overshadows it.

Future work includes obtaining more data that could be used to train the models, either from alternative forums or platforms, or from the different applications that were used during this period. Alternative methods for prediction and the inclusion of features derived from alternative manipulation of the data like sentiment analysis could lead to new insights that better explain the phenomenon and help generate an acceptable model.

References

[1] Provost, F., Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking (1st ed.). O'Reilly Media. https://www.researchgate.net/publication/256438799_Data_Science_for_Business

[2] M. Yearworth and L. White, "Spontaneous emergence of Community OR: Self-initiating, self-organising problem structuring mediated by social media," vol. 268, no. 3, pp. 809–824, Aug. 2018, doi: 10.1016/j.ejor.2018.01.024.

[3] C. Leong, I. Faik, F. T. . Tan, B. Tan, and Y. H. Khoo, "Digital organizing of a global social movement: From connective to collective action," vol. 30, no. 4, Dec. 2020, doi: 10.1016/j.infoandorg.2020.100324.

[4] Chen, W., Zhang, H., Mehlawat, M. K., Jia, L. (2021). Mean–variance portfolio optimization using machine learning-based stock price prediction. Applied Soft Computing, 100. https://doi.org/10.1016/j.asoc.2020.106943

[5] Jing, N., Wu, Z., Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications, 178, 115019. https://doi.org/10.1016/j.eswa.2021.115019

[6] Li, X., Wu, P., Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing Management, 57(5), 102212. https://doi.org/10.1016/j.ipm.2020.102212

[7] Hitam, N. A., Ismail, A. R., Saeed, F. (2019). An Optimized Support Vector Machine (SVM) based on Particle Swarm Optimization (PSO) for Cryptocurrency Forecasting. Procedia Computer Science, 163, 427–433. https://doi.org/10.1016/j.procs.2019.12.125

[8] Chao, L., Zhipeng, J., Yuanjie, Z. (2019). A novel reconstructed training-set SVM with roulette cooperative coevolution for financial time series classification. Expert Systems with Applications, 123, 283–298. https://doi.org/10.1016/j.eswa.2019.01.022

[9] Zhou, Z., Gao, M., Liu, Q., Xiao, H. (2020). Forecasting stock price movements with multiple data sources: Evidence from stock market in China. Physica A: Statistical Mechanics and Its Applications, 542, 123389. https://doi.org/10.1016/j.physa.2019.123389

[10] Rezaei, H., Faaljou, H., Mansourfar, G. (2021). Stock price prediction using deep learning and frequency decomposition. Expert Systems with Applications, 169, 114332. https://doi.org/10.1016/j.eswa.2020.114332

[11] J, K., Sengupta, I., Chaudhury, S. (2018). Stock Market Prediction Using Time Series Analysis. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3168423

[12] Ji, L., Zou, Y., He, K., Zhu, B. (2019). Carbon futures price forecasting based with ARIMA-CNN-LSTM model. Procedia Computer Science, 162, 33–38. https://doi.org/10.1016/j.procs.2019.11.254