NLP in the Financial Market

Modeling stock price returns by using sentiment indicators derived from financial news

Project Advisor: Thomas J. Diciccio

Junyue Wang (jw2252)

Jinqi Song (js2884)

Zixuan Wang(zw352)

Fanfei Gu(fg295)

Jayaram Gokulan(jg929)

May 12, 2022

A paper presented for the MPS Project.

Cornell Bowers CIS
Statistics and Data Science

Acknowledgement

This project is sponsored by Gravity Investments. We would like to thank Mr. James Damschroder for giving us project goals and valuable advises for this project. Moreover, the guidance provided by Professor Thomas J. Diciccio was much appreciated.

Contents

Executive Summary

Overview

Stock market and its trends are extremely volatile and hard to predict. However, with real-time information available to us on the Internet, there are plenty of unstructured text data that can capture some of their volatility and be used for predicting next moves. Moreover, some study shows stock market movements can be influenced by information shared through news, and social media (Fataliyev et al., 2021; Mohan et al., 2019).

Problem Statement

In this project, we want to develop a model that captures sentiment data from original news sources. As used in many literature, we uses mainly python to extract those information (Boh, 2022). Then we test and research the data to tune it into a useful predictive signal for stock price changes.

Our Conclusion

- Web crawlers and techniques such as random forest, fixed/mixed effect models, VAR models are used in our analysis.
- Sentiment alone is not a strong signal to model the whole market change. Some of its derivatives (e.g. negative news count/day) on the contrary are more robust in terms of stock price.
- The VAR model fitted on the sentiment and stock price time series produced good prediction results for the stock prices when time lagged sentiments are used as predictors.

Final Thoughts and Next Steps

This paper adds to different literature by using various programming and statistical techniques on creating predictive sentiment features for market changes. In all, the sentiment features we've created can be a useful predictive signal for stock price changes. However, the limitation of our analysis mainly rises because of the unavailability of news data for a larger time span. More work would be done in data collecting and handling procedures for larger time periods.

1 Introduction

In the financial field, with the advent of the era of big data, the speed of information production and transmission is unprecedented. We have loads of textual information booming every day in social media. For instance, we have finance news, company annual reports, company announcements and analyst reports from a more formal point of view. We also have social media where people are sharing how they think of the stock market, such as twitter and reddit from a more casual point of view. These are all useful text sources to be potentially analyzed and draw conclusions. However, the text data often lacks effective analysis tools. In the past, the analysis efficiency has been relatively low while the cost of analysis has been high. Besides, textual data is generally more challenging to draw information from compared to numerical data due to the nature of text, especially for computers. So there is an escalating demand for more efficient analysis tools that can quickly skim through those texts and pick up useful information. Nowadays as technologies in deep learning are becoming more and more mature, NLP(Natural Language Processing) can provide a solution to deal with financial text sentiment analysis.

Over the past few years the interest in sentiment analysis has grown rapidly and it is believed that the interest in sentiment analysis will still be tremendous in the next decade since market sentiment is one factor that can influence the trend of the stock market. Market sentiment refers to the overall attitudes of investors towards the financial market or a specific stock. It is a combination of facts and rumors. It becomes bullish when stock prices are rising and it becomes bearish when stock prices are falling. Investors often rely on market sentiment before they make any trading actions because market sentiment is usually an indicator of the trend of the stock market. For example, in March 2020, due to the global pandemic, the ISEE sentiment index, a measurement of investor sentiment in the market, dropped nearly from 70 to 40. At the same time, the SP 500 index dropped almost from 3300 to 2500. This is also known as the 2020 Coronavirus Stock Market Crash, which occurred when people panic selling their stocks. This example shows the relationship between market sentiment and stock price returns.

Sentiment analysis is a useful text mining method to process textual content and meaningful information. Market participants have recently raised growing interest in the field of sentiment analysis.

With the development of algorithmic trading systems human-decision making has reached its limit and cannot keep up with processing and execution of orders decisions. (Uhr et al., 2014) This is why sentiment analysis is very needed. It can make judgments about the positive or negative nature of text or sentences. What this paper focuses on is sentiment analysis of the financial market. Through sentiment analysis, hundreds of news articles about listed companies can be analyzed quickly and given a sentiment score ranging from -1 to 1, where -1 represents negative sentiment and 1 represents positive sentiment and 0 represents indifferent sentiment. Then, these sentiment scores can be used to possibly predict future stock price changes and provide useful insights for taking future stock trading actions.

The rest of this study is organized as follows. First, the data generating processes of stock prices and news and the dictionary are discussed in section 3. We have our stock price data from tiingo and news sources from yahoo finance. We have our own dictionary by combining VADER and NTUSD. Section 4 discusses the data preprocessing techniques, such as matching the stock price data and news, generating some additional data features of market sentiment and using some machine learning techniques to find some significant features. Section 5 gives a walk-through of the models that we will apply in section 6. Section 6 discusses fitting the fixed effect model, the mixed effect model and the multivariate time series model in great detail. Section 7 presents the results of the models from section 6 and gives some inferences. Finally section 8 concludes the paper and gives some final discussions.

2 Problem Statement

Currently with the amount of textual data that exist, how to make a meaning useful of textual sources has been a challenge. In the financial market, in particular, news is information about companies and there is news that gets published every second. Out of many textual resources, news is considered to be relatively objective and thus can generalize the current market sentiment. In this project, we try to implement a stock news sentiment analysis model that helps to overcome the challenges of identifying the sentiment of financial news and turn these sentiments into useful signals to predict stock price changes. Several techniques have been utilized in this project, such as web scrap-

ing,feature engineering,statistical modeling and forecasting. By using these techniques, we successfully generated,cleaned and modeled our data. The conclusion we get is that sentiment score alone is not a strong predictor of stock price changes. However when we model the data using the VAR model, we can get good forecasts.

3 Data and Description

This project uses data for for about 110 stocks. This includes the NASDAQ 100 stocks and some stocks that the client are interested in and another ten as a supplement. The selection of news is based on data-availability and covers stocks from a range of price levels. The main independent variable is sentiment scores, which are later modified to other sentiment features explained in section 4.

3.1 Price and News Retrieval

For stock prices, we uses Tiingo API to download daily adjusted closing stock prices as one of our data sources. This is also consistent with client's company resources .

We use news from Yahoo Finance to retrieve headlines and contents. We store the news headline and contents separately and then calculate the sentiment score for each of them. Due to the nature of the Yahoo Finance website, the number of news per day is different for every stock, and the time span for news are different for each stock. The exact process that we used to parse and filter the news is explained in section 4.

3.2 Sentiment Score Dictionary

After the news are being parsed and scripted, we create a dictionary to hold the sentiment score for words that might appear in the news (part of the dictionary shown in figure 1 and figure 2 in the Appendix). Then we use the dictionary to calculated the sentiment scores for the texts that we get from the news.

We use a combination of two dictionaries. The first one is VADER, which stands for the Valence Aware Dictionary and Sentiment Reasoner (Hutto and Gilbert, 2014). The second one is NTUSD-fin, which is part of the National Taiwan University Semantic Dictionary (Chen et al., 2018). Our final

dictionary is based on those two dictionaries, in addition to some modifications and some of our chosen words.

3.2.1 VADER

The VADER in the NTLK package in Python contains mix of words that are marked by their semantic direction, positive or negative. It shows both the polarity and the intensity of sentiments expressed in social media micro-blogs.The words are developed from common sentiment expressions and contains about 8000 words with grammar and punctuation. This dictionary is more human-based: when creating the dictionary, individual human raters rate the words from extremely negative to extremely positive (Hutto and Gilbert, 2014).

To make them more applicable to our case, we add a few other words that might show up on financial news such as impressed, blasted, abandoned, high risk, etc.

3.2.2 NTUSD-fin

This package is generated using ML methods and include words that might not normally comes into mind when thinking about sentiment expressions and it is more finance-related. It uses StockTwits as the data source to collect the words. StockTwits is a Twitter-like social media for investors to share their information and opinions of the market or a certain company (Batra and Daudpota, 2018). When creating posts, the bullish and bearish bottoms allow users to label their market sentiment in the post. The sentiment score is calculated by the correlation between words and sentiment (Chen et al., 2018).

All the two dictionaries are adjusted to the same scale in order to be used together. We use similar calculation as in the VADER package but added the words in the NTUSD dictionary. The final sentiment score of a text is calculated by normalizing the sum of all sentiment of each individual word. An example of the news and its sentiment score is shown in Figure 4.

Before we do any data-processing, We draw a dual-axis plot for stock price and sentiment score for news title and find out that their overall trends are somewhat similar and can be further analyzed (an example is shown in figure 5). Some modification of the sentiment score data and the research on their relationships are shown in section 4 and 5.

4 Data Preprocessing

In this part, we performed preprocessing to organize messy data. This contains removing company with less than 30 days of news, rematching news time and stock time, excluding less informed news and new feature engineering.

4.1 Data overview

Suggested by the subject of our project, the main goal it to extract useful information from pertinent news and articles and analyse how they affect stock price accordingly. We managed to parse news from Yahoo finance by Ajax loading, digging into web sources and building our own scrawler to get title, article text, date, link and etc. Ajax loading uses a combination of: A browser built-in XMLHttpRequest object (to request data from a web server) / JavaScript and HTML DOM (to display or use the data). Parsing from an open source is extremely hard, especially here we encountered many technical difficulties, such as some web page requires authentication or login to other pages. We cannot redirect to other sources as we fail to discover a general format in HTML code. We mainly focus on yahoo original sources, and other sources that were embedded into yahoo webpage frame. The news data contains 17653 observation over 110 companies.

4.2 Scores Distribution

As we are trying to find realtion between text and stock, we focus on two variables: title and article text. Ideally, title is a summary of article and it contains major, but incomplete information. Article content means to provide more in-depth information and details the signal that may lead to increase or decrease in price. When we looked into the ditribution, we found text score extremely skewed, with points clustering over high value close to 1. However, title scores display a close normal distribution centering around 0.5, which is ideal for analysis even further transformation. (Figure 3) While we doubted this scenario where article is overall overblown, we read through multiple news from different companies and found a common case: News are generally inflated and biased towards readers. Even for a negative report, it could use many turning words like but, however followed by other cases that company is making positive signs. We decided to stick with title score for analysis given its bal-

anced distribution and interpretability.

4.3 Feature Engineering

So far, we have had cleaned data and sentiment scores ready for analysis and model building. According to customer's requirement and our research, we generated the following new features. Besides, to see their effects towards stock market in a long run, we modified original format with moving average. Limited to the size of data and resources, we can only pick shorter time range, that is a window size of 5/10/15/20 days. On top of this, we can have general idea on how these variables perform over stock price and the insight into possibility of longer time period for moving average strategy.

- Volume weight
- Positive count
- Negative count
- Net sentiment (revised sentiment scores by updating library)
- Net sentiment change (revised sentiment change)
- Net sentiment change observed only when sentiment direction is opposite of periodic price change
- Article count (number of news per day)
- Article Count * Net sentiment (intercation term)

Volume weight refers to popularity and number of days for news resources. If one company has smaller value for volume weight, this company has more popularity among investors or media. Positive Count/ Negative count are number of positive/negative news per day based on 0.05 sentiment score. Net Sentiment is new sentiment score calculated through our revised sentiment library and formula, where we change max capacity from 15 to 20 to hold more extreme values. Variance is expected to be smaller that previous scores. Net Sentiment Change is daily value difference grouped by company, where we were really careful to avoid a misallocation of different companies. Article count

is daily number of news. Article count * Net sentiment works as an interaction term to determine potential effects by considering a factor that net sentiment change is expected to be more flat if one company has large volume of news per day. All these variables have significant financial meanings and are usually taken into consideration in more advanced NLP model. Next, we try to figure out which ones of these variables closely pertinent to stock price in our cases.

4.4 Feature Selection

As stated above, we have some new features added in out sentiment analysis. Clearly, the moving average ones are highly correlated, because they were transformed from original numerical value. Therefore, we ran feature selection by machine learning algorithm in this step to pick most influential ones. Two models with variable shrinking/ranking nature were used here: Random Forest and Lasso. Random Forest returns feature importance by decreasing portion of Gini index (purity). If we include one important variables, we could expect a large decrease in variance of final results. For Lasso model, it adds regularization term to penalize variable size. We use ten folds cross validation method to pick best parameter from 0.1 to 10 by 0.1 with evaluation metrics R^2. Since we are fitting a regression problem here, R^2 measures how much this model would explain variance in dataset. Finally, Lasso model returned a sparse matrix with 0 indicating variables showing non-significance and coefficients show relative importance (larger coefficient means more importance). Comparing these two results, Random Forest returned 5 variables and Lasso model returns 3 variables (with large regularization term).

Usually, large variable size would lead to a lower training error as the model flexibility increases with number of predictors, but it would lead to overfitting problem, and thus dampening test result. In our cases, the prediction error could be lower in train set, but we worried it would not work as expected on outer data. To retain both efficiency and conciseness, we took the intersection of these two sets. Finally, we have three variables out of 28, which are count / Neg count 20 / net sentiment 15. From the result, we can tell that these variables are useful particularly in long term. This conforms with fact that most news does not have immediate impact on stock price until one or two days later.

5 Empirical Specification and Models

In this section we summarize the models we've used in this study and its result. The result tables are shown in the Appendix.

5.1 Fixed Effect and Mixed Effect Regression

5.1.1 Fixed Effect Regression

The entity-demeaned fixed effects regression result is shown in Figure 7. The independent variable includes demeaned count, demeaned negative sentiment change with 20 days moving average, and demeaned net sentiment change with 15 days moving average. The dependent variable is demeaned stock price. The p-values are all less than 0.01, which indicates a significant relationship between stock price and sentiment score. The count variable does not really matter here, because the coefficient is nearly 0. Regarding coefficient of demeaned negative sentiment change, we can interpret it as one unit increase in negative sentiment change will lead to 16 units decrease in stock price. This large-unit price drop makes sense because negative news always have devastating effects. People may not choose to buy in when hearing some positive news, but definitely will stop buying when hearing negative news. Similarly, one unit increase in demeaned net sentiment change will make stock price reduce 8 units, which also makes sense since fluctuations in sentiment scores will diminish a stock's credit and people will be reluctant to make purchases.

5.1.2 Mixed Effect Regression

As we can see from Figure 8, mixed effects have similar results as fixed effects. Therefore it is reasonable to conclude here that our data might not contain cross effects part.

5.2 Vector Autoregression & Vector AutoRegression Moving Average Model

5.2.1 Preliminary test results for Multivariate Time series models

a. Correlation tests :

Correlations were examined between lagged values of sentiments and stock prices. The results obtained served as a prima facie evidence to assert the presence of a linear relationship between stocks and sentiments.

Results : A Pearson correlation test conducted between sentiments and stock prices showed statistical significance for certain lagged values at the 90 percent confidence level shown in figure 11 and 12. The columns in this figure represents lag values while the rows depicts individual stocks.

b. Granger Causality Tests :

The primary interest of this project is to learn whether sentiment scores, derived from section x, could be a significant predictor for stock prices. Establishing a direction for the said relationship determined by correlation tests is thus paramount. Granger Causality test helps identify the direction of causation between any two data series by examining the sufficiency of one time series data in predicting the other. As an adequate indicator for determining the causal variables, our research employed the Granger causality hypothesis test on the sentiment and stock price data to give the results listed in figure 9 (*How to Perform a Granger-Causality Test in Python*, n.d.).

Results : A chi squared test conducted at a 10 percent significance level helped confirm that only 42 stocks from the portfolio could effectively have sentiments as significant predictor in its forecasts of stock prices.

c. Stationarity and Co-integration Tests :

Time invariant statistical properties are central to make consistent, accurate predictions and forecast from a given time series. This study implemented the Augmented Dickey Fuller Tests to evaluate the stationarity of the both sentiments and stock prices time series (*Cointegration tests on time series*, n.d.).

Correlations that were investigated in the prior section can be extended to cover a longer time frame. This phenomenon, termed co-integration in the world of time series, refers to the situation when the two time series data are integrated i.e have a non deviating spread for a long time frame. Our study investigated the cointegration of sentiments and stock prices by measuring the stationarity

of the spread between both these series. This study employed the Augmented Dickey Fuller (ADF) Test to evaluate the stationarity and cointegration of stock price and sentiment time series data.

Results : As seen from the Figure 10, the stocks COST and LPL were not sentiment stationary as per ADF tests, while MRNA was not stock price stationary using the Augmented Dickey Fuller test. Also every stock apart from the COST, MAR, MRNA QCOM had their price and sentiments co-integrated

5.2.2 General Specifications and Procedures for VAR/VARMA model fitting

This study successfully fitted both a VAR and a VARMA model to the final list of stocks, allowing for a maximum of 10 lags, and considered two criteria namely AIC and Root Mean Squared Error(RMSE) as benchmarks to evaluate forecast accuracy of these models. It is to be noted that the VAR was generated by specifying a maximum lag of 5 while VARMA model had several possible orders for the AR and MA components evaluated, ranging from 0 to 5. The orders giving the lowest value for AIC and RMSE are specifically noted. Additional effort was made to analyze the forecast efficacy of VAR/VARMA model on a single random stock from our portfolio.

5.2.3 VAR Model Results

The best VAR model (order) for our portfolio of stocks are illustrated in the figure 14. These results have been collated considering the best lags that AIC and RMSE allots for the respective stocks. Often the RMSE implied lags matched the values suggested by the AIC criteria, but overall there wasn't a clear consensus on the best lags for the model taking either of these parameters into consideration.

5.2.4 VARMA Model Results

The ideal VARMA model order shown in figure 13 involved evaluating several candidate orders for AR and MA component of the series and selecting those with lowest AIC and RMSE values for the corresponding model. As can be seen from these results often times the AIC criteria suggested a more parsimonious model compared to RMSE, while delivering less accurate out of sample forecasts.

5.2.5 Model Evaluation considering return on portfolio

The inferences drawn from our study's results vary depending on the user requirement. Since more than one stock was considered in this study, one parameter of interest is the return on the entire portfolio of stocks. Selecting the best model (VAR/VARMA) in terms forecast performance becomes an immediate goal. RMSE values are subsequently scrutinized across either models for every stock and models with lower RMSE become the go to model for future investment decisions on such stocks. The enhanced prediction accuracy on the portfolio optimizes the return as per investor preference.

5.2.6 Model Evaluation considering individual stocks

An alternative way to evaluate the model, is based on individual stock performance. Comparing the RMSE generated by VAR/VARMA for the same stock informs the user of the model's prediction power. Among the stocks that VARMA model favoured include those from the manufacturing and health care industry. Although this may not constitute compelling evidence to suggest that VARMA models are more suitable for such stocks, it certainly does shed light on revealing sectors where lagged values of sentiments and their error terms can be significant predictors.

As an illustration, this study chose to evaluate the CPRT stock using both VAR and VARMA model (best order) giving the coefficient estimates seen in the statistical summary shown in figures from 16 to 19.

1. VAR gave the following coefficients, based on a 5 percent significance level.

$$StockPrice_t = -0.01 + 0.02StockPrice_{t-1} + 0.004Sentiment_{t-2}$$
$$- 0.007Sentiment_{t-3} - 0.02Sentiment_{t-4} - 0.02Sentiment_{t-5} - 0.01StockPrice_{t-5} + \epsilon_t$$

2. VARMA model, on the other hand, produced the below coefficients, based on a 5 percent signif-

icance level.

$$StockPrice_t = -0.78\epsilon_{t-1} \tag{1}$$

Thus VARMA model didn't imply any lagged value of sentiment or its errors as a significant enough predictor for stock prices. Furthermore, it was observed that RMSE favoured the VAR model generating a much lower value compared to the VARMA model. A similar investigation on the remaining stocks of our portfolio provide some level evidence to assert that on an average VAR models outperformed their VARMA counterparts when it came to prediction accuracy. However, without further investigation it still remains to be seen how well this can be generalized to every other stock within the same industrial category or any other classification for that matter.

6 Theory

Here we give a brief sketch and walk-through of the modeling theories and how that is applied to price and sentiment data.

6.1 Fixed Effect and Mixed Effect Regression

In order to better explain the data, we use several models to explore the relationship between sentiment scores and stock prices.

6.1.1 Fixed Effects

The first model we incorporate is fixed effects model. Fixed effects model is a statistical method in which the intercept of regression model are allowed to vary freely across groups. Fixed effects are variables that are constant within group, which don't change or change at a constant rate over time. With these across-group variations controlled, individual-specific variances can be better exploited and understood.

This model can be applied to our data because each company has their own company-specific attributes. So, multiple dummy variables are generated in the stock data. The formula we use can be

expressed as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_2 + \gamma_3 D_3 + ... + \gamma_n D_n + \mu_{it}$$

However, one problem with this model is that when groups go large, the regression can become tedious. In our case, we will have a regression result with 101 coefficients. To resolve this problem and make the results more efficient, we use entity-demeaned fixed effects model instead. Entity-demeaned means subtracting each observation by its mean value to rule out entity unique, but group invariant impact from outcome variable. The formula can be expressed as

$$Y_{it} - \bar{Y}_{it} = \beta_1 (X_{it} - \bar{X}_{it}) + (\mu_{it} - \bar{\mu}_{it})$$

(Chen, 2021).

6.1.2 Mixed Effects

Besides the fixed effects that are root to each company, we wondered if there exists some random effects that all company share together? That is saying, one variable has different scale of effect on each company, but we could further process its mean effects and apply this amount effect on other unseen companies and their stock. In this case, we were trying to generalize the mean effects that randomly take apart across companies. The random part actually captures the covariates in design matrix. The key assumption this model lies in is the design matrix is not independent. It is possible in our case that one company is a spin-out company of another, which means a corporate realignment involving separation of a division to form new corporation. The model is expressed as:

$$y1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + a1 || y2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + a2$$

where a1 and a2 are correlated here, instead of independent for normal linear regression. (Figure 6) (*PYTHON FOR DATA SCIENCE*, n.d.).

6.2 Vector Autoregression & Vector Autoregression Moving Average Model

6.2.1 Introduction

The multilevel modelling approach discussed in section 5.1 evokes the more general stochastic regression and state space models used in time series modelling. State space models are an effective medium to understand observations resulting from a time dependent causal factor. Observing the sentiment and stock price series presented in this study, there was reason to believe that an underlying common state process could have generated either of the series'. A multivariate times series model was thus recommended to capture the interaction between sentiments and stock prices.

A Vector Auto Regression is a structural framework that encapsulates and models such interactions between two or more time series vectors. In an effort to understand the interplay between times series, this study proposed to develop two kinds of multivariate time series models; the Vector Auto Regression (VAR) and the more general Vector Auto Regression Moving Average (VARMA) model. While VAR assumes that the times series' own lags can effect its future values, VARMA posits that lags of the white noise errors terms, from both times series, can be an additional influence.(Lütkepohl, 2005).

A 1st order VAR model is depicted below.

$$StockPrice_t = \alpha_1 + \beta_1 StockPrice_{t-1} + \beta_2 Sentiment_{t-1} + \epsilon_t \tag{2}$$

$$SentimentPrice_t = \gamma_1 + \delta_1 Sentiment_{t-1} + \delta_2 StockPrice_{t-1} + \omega_t \tag{3}$$

where α_1 denotes constant term for mean adjusted stock price VAR model

β_1 denotes coefficient for the first time lagged stock price values in stock price VAR model

β_2 denotes coefficient for the first time lagged sentiment values in stock price VAR model

ϵ denotes white noise term for the stock price VAR model

γ_1 denotes constant term for mean adjusted sentiment price VAR model

δ_1 denotes coefficient for the first time lagged sentiment values in sentiment VAR model

δ_2 denotes coefficient for the first time lagged stock price values in sentiment VAR model

ω denotes white noise term for the sentiment VAR model

A 1st order VARMA model is similar to the VAR model, subject to the addition of moving average terms as shown below.

$$StockPrice_t = \alpha_1 + \beta_1 StockPrice_{t-1} + \beta_2 Sentiment_{t-1} + \beta_3 \epsilon_{t-1} + \beta_4 \omega_{t-1}$$

$$SentimentPrice_t = \gamma_1 + \delta_1 Sentiment_{t-1} + \delta_2 StockPrice_{t-1} + \delta_3 \omega_{t-1} + \delta_4 \epsilon_{t-1}$$

where α_1 denotes constant term for mean adjusted stock price VAR model

β_1 denotes coefficient for the first time lagged stock price values in stock price VAR model

β_2 denotes coefficient for the first time lagged sentiment values in stock price VAR model

β_3 denotes coefficient for the first time lagged error term of stock price VAR model

β_4 denotes coefficient for the first time lagged error term of sentiment VAR model

ϵ denotes white noise term for the stock price VAR model

γ_1 denotes constant term for mean adjusted sentiment price VAR model

δ_1 denotes coefficient for the first time lagged sentiment values in sentiment VAR model

δ_2 denotes coefficient for the first time lagged stock price values in sentiment VAR model

δ_3 denotes coefficient for the first time lagged error term of sentiment VAR model

δ_4 denotes coefficient for the first time lagged error term of stock price VAR model

ω denotes white noise term for the sentiment VAR model

7 Conclusion and Discussion

Our project meant to discover key factors that contribute to explanation and forecasting of stock price. According to analysis, sentiment score itself does not have enough predictive power to cover variations in market. Some new features and their moving averages do contribute to modeling of stock price. Next, we used fixed effects and random effects model to verify and quantify their scale of

effects towards target. The vector autoregression model and its variant, vector autoregression moving average model implemented in this study manages to uncover sentiments role as time dependent causal factor in stock price prediction. However, the results are based on limited number of stocks and their dependence with sentiments. How well our results generalizes to a random stock still needs to be investigated on, before arriving at any solid conclusions. Care should be taken to note that the results mentioned here are based on a 5 - 10 percent significance level. This is subject to change depending on an investors risk profile and preference.

Our analysis may not be perfect here, but it serves as a good starting point for industrial use of sentiment score and its extensions. First, the resources are limited. We are restricted to technical solution of getting more sentiment text either from social media or other platform. That would require a recursive updating strategy, preset cleaning pipelines and huge database to store the data. By incorporating larger amount of data, longer time range analysis and moving average is feasible, and, suggested by our analysis, variables with longer time period may posses power in terms of modeling price change. Second, the way we calculated sentiment score now is searching for single word and matching with values in our dictionary. Other methods like N-gram and TF-IDF could be applied here for an accurate measurement. Third, moving on from our analysis, a stratified prediction grouped by time periods could be formed for actual return combining these new features.

Sentiment analysis application in financial world is trending now. We sincerely wish to facilitate research in this area to a new level with precise analytical results and well tuned mechanisms.

Appendix

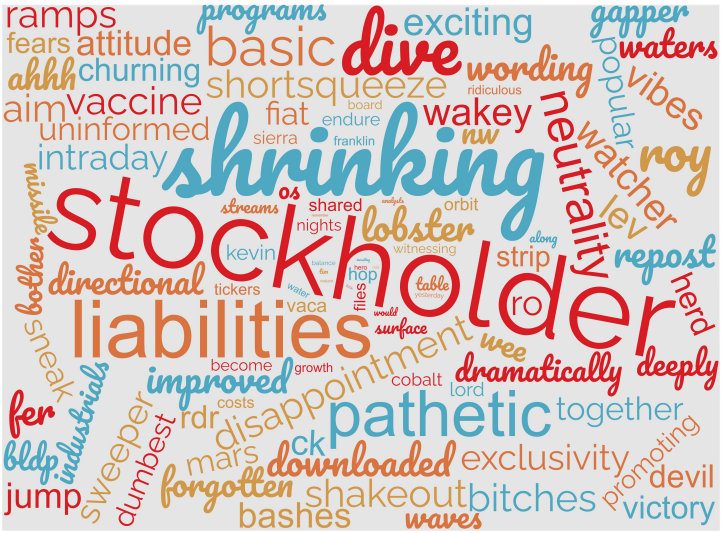


Figure 1: Part of the words in our sentiment dictionary

```
'abandon': -3.296990464331529,
'abandoned': -1,
'abandoner': -1.9,
'abandoners': -1.9,
'abandoning': -1.6,
'abandonment': -2.4,
'abandonments': -1.7,
'abandons': -1.3,
'abducted': -2.3,
'abduction': -2.8,
'abductions': -2.0,
'abhor': -2.0,
'abhorred': -2.4,
'abhorrent': -3.1,
'abhors': -2.9,
'abilities': 1.0,
'ability': -1.0373304531733,
```

Figure 2: A snapshot of final dictionary with word:score

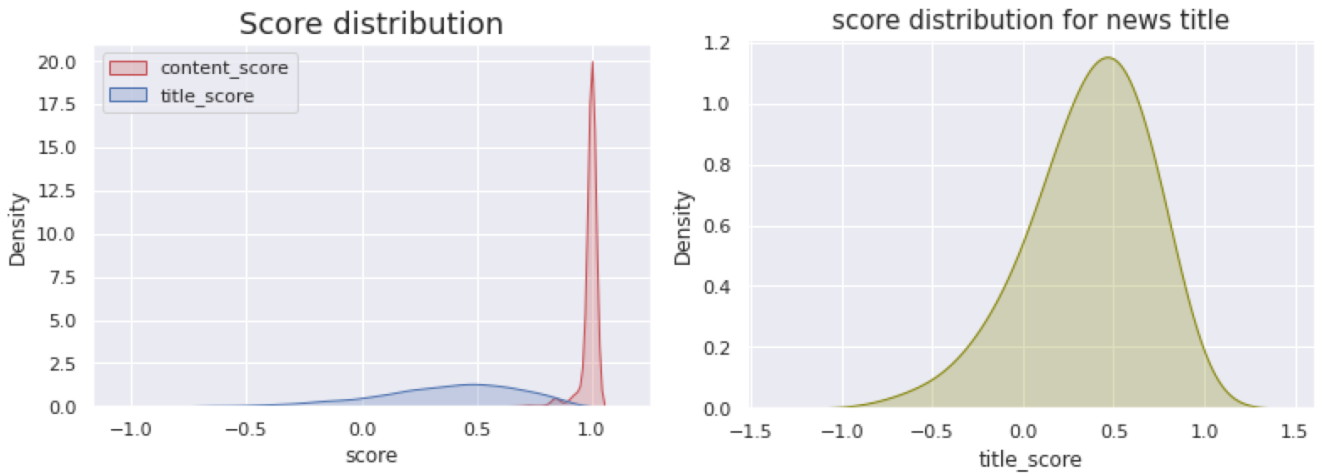


Figure 3: Sentiment Score Distribution

```
Apple Stock Called 'Good Hiding Place' In Volatile Market
Unions are on the rise. Guess why.
Activision Union Bid Will Go to a Vote, Labor Board Orders
Foxconn’s Key iPhone Plant Operating in Locked-Down China Region
Top 5 Things to Watch in Markets in the Week Ahead
```

	neg	neu	pos	compound
1	0.238	0.064	0.697	0.6280
2	0.000	0.833	0.167	0.0525
3	0.097	0.131	0.771	0.7401
4	0.243	0.301	0.456	0.4266
5	0.372	0.268	0.360	-0.4899

Figure 4: Example of News Title and its Sentiment Score

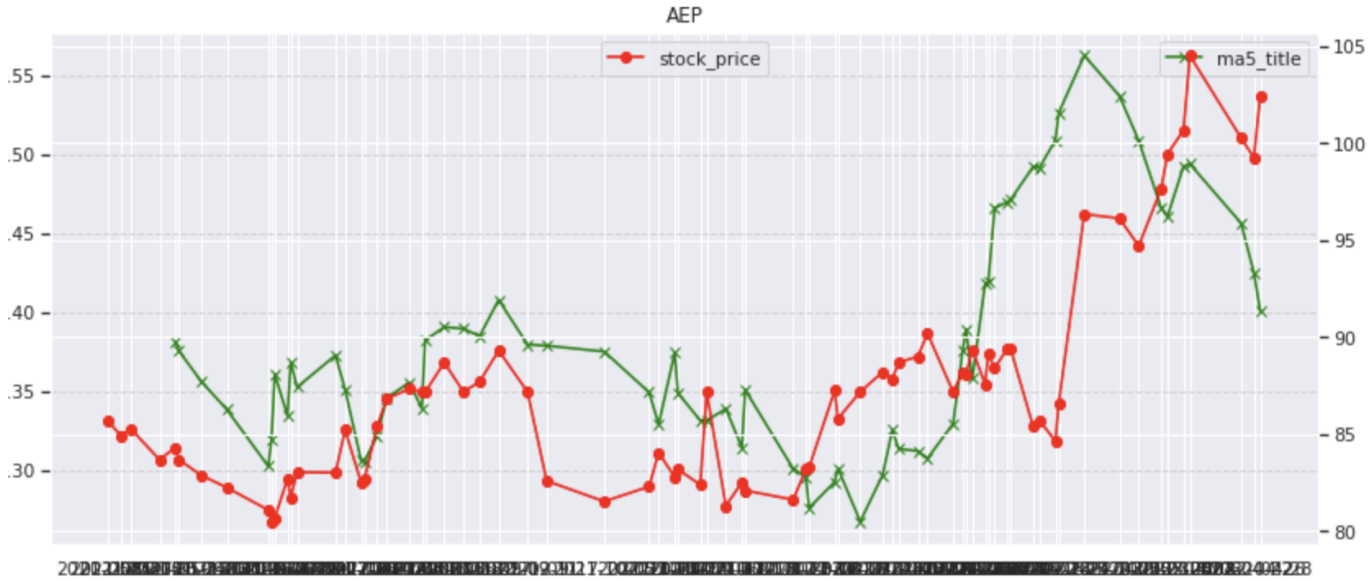
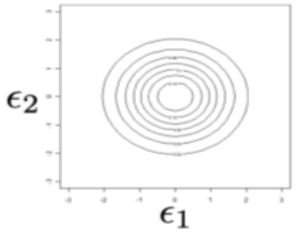


Figure 5: Compare AEP stock prices and sentiment scores using dual-axis plot

To help understanding

- For example, for *n*=2:

$$y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + \epsilon_1$$
$$y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + \epsilon_2$$



- What if we introduced a correlation?

$$y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + a_1$$
$$y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + a_2$$

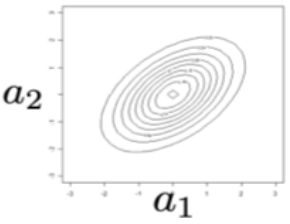


Figure 6: Mixed effects

OLS Regression Results

Dep. Variable:	new_price	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	23.30
Date:	Mon, 02 May 2022	Prob (F-statistic):	8.19e-11
Time:	18:58:38	Log-Likelihood:	-36294.
No. Observations:	7407	AIC:	7.259e+04
Df Residuals:	7404	BIC:	7.262e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.803e-15	0.378	1.01e-14	1.000	-0.740	0.740
new_count	4.254e-15	6.23e-16	6.825	0.000	3.03e-15	5.48e-15
new_neg_20	-16.1476	2.444	-6.608	0.000	-20.938	-11.358
new_net_sentiment_15	-8.6375	3.223	-2.680	0.007	-14.955	-2.320

Omnibus:	1920.227	Durbin-Watson:	0.546
Prob(Omnibus):	0.000	Jarque-Bera (JB):	144335.576
Skew:	-0.198	Prob(JB):	0.00
Kurtosis:	24.622	Cond. No.	1.48e+18

Figure 7: Fixed Effects Regression Results

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	stock_price
No. Observations:	7407	Method:	REML
No. Groups:	97	Scale:	1070.2816
Min. group size:	30	Likelihood:	-36752.4271
Max. group size:	123	Converged:	Yes
Mean group size:	76.4		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	166.403	103.730	1.604	0.109	-36.905	369.711
count	1.082	1.285	0.842	0.400	-1.437	3.601
neg_count_20	-16.147	2.459	-6.565	0.000	-20.967	-11.326
net_sentiment_15	-8.620	3.244	-2.657	0.008	-14.979	-2.262
stock_code Var	83452.909	371.051				

Figure 8: Mixed Effects Regression Results

	Cause	Effect	Lags
ADP	diff_stock_price_x	diff_title_score_y	1.000000
ALGN	diff_title_score_x	diff_stock_price_y	3.000000
AVGO	diff_title_score_x	diff_stock_price_y	1.000000
BIDU	diff_stock_price_x	diff_title_score_y	3.000000
CHTR	diff_stock_price_x	diff_title_score_y	3.000000
CME	diff_stock_price_x	diff_title_score_y	2.000000
COIN	diff_title_score_x	diff_stock_price_y	3.000000
COST	diff_title_score_x	diff_stock_price_y	3.000000
CPRT	diff_stock_price_x	diff_title_score_y	1.000000
CSX	diff_title_score_x	diff_stock_price_y	2.000000
EBAY	diff_stock_price_x	diff_title_score_y	2.000000
EXC	diff_title_score_x	diff_stock_price_y	3.000000
FISV	diff_title_score_x	diff_stock_price_y	1.000000
HON	diff_title_score_x	diff_stock_price_y	1.000000
IDXX	diff_stock_price_x	diff_title_score_y	1.000000
ILMN	diff_stock_price_x	diff_title_score_y	2.000000
ISRG	diff_title_score_x	diff_stock_price_y	3.000000
KDP	diff_stock_price_x	diff_title_score_y	3.000000
KHC	diff_title_score_x	diff_stock_price_y	1.000000
LCID	diff_title_score_x	diff_stock_price_y	1.000000
LPL	diff_stock_price_x	diff_title_score_y	3.000000
LRCX	diff_stock_price_x	diff_title_score_y	2.000000
LULU	diff_stock_price_x	diff_title_score_y	3.000000
MAR	diff_title_score_x	diff_stock_price_y	3.000000
MRNA	diff_stock_price_x	diff_title_score_y	3.000000
MRVL	diff_title_score_x	diff_stock_price_y	1.000000
MU	diff_stock_price_x	diff_title_score_y	1.000000
NTRS	diff_title_score_x	diff_stock_price_y	2.000000
PAYX	diff_title_score_x	diff_stock_price_y	1.000000
PCAR	diff_title_score_x	diff_stock_price_y	3.000000
PDD	diff_title_score_x	diff_stock_price_y	1.000000
PEP	diff_title_score_x	diff_stock_price_y	2.000000
QCOM	diff_stock_price_x	diff_title_score_y	2.000000
SNPS	diff_stock_price_x	diff_title_score_y	2.000000
TEAM	diff_title_score_x	diff_stock_price_y	1.000000
TMUS	diff_stock_price_x	diff_title_score_y	3.000000
TXN	diff_stock_price_x	diff_title_score_y	3.000000
VALE	diff_title_score_x	diff_stock_price_y	2.000000
VRSN	diff_stock_price_x	diff_title_score_y	3.000000
WDAY	diff_stock_price_x	diff_title_score_y	2.000000
XEL	diff_stock_price_x	diff_title_score_y	3.000000
ZS	diff_title_score_x	diff_stock_price_y	1.000000

Figure 9: Granger Causality Test Results

	Sentiment Stationarity	Stock Price Stationarity	Cointegrated
ADP	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
ALGN	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
AVGO	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
BIDU	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
CHTR	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
CME	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
COIN	Yes - p value 0.02	Yes - p value 0.01	Yes - p value 0.01
COST	No	Yes - p value 0.0	No
CPRT	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
CSX	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
EBAY	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
EXC	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
FISV	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
HON	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
IDXX	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
ILMN	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
ISRG	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
KDP	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
KHC	Yes - p value 0.01	Yes - p value 0.0	Yes - p value 0.01
LCID	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
LPL	No	Yes - p value 0.0	Yes - p value 0.0
LRCX	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
LULU	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
MAR	Yes - p value 0.04	Yes - p value 0.0	No
MRNA	Yes - p value 0.0	No	No
MRVL	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
MU	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
NTRS	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
PAYX	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
PCAR	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
PDD	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
PEP	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
QCOM	Yes - p value 0.0	Yes - p value 0.0	No
SNPS	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
TEAM	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
TMUS	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
TXN	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
VALE	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.01
VRSN	Yes - p value 0.0	Yes - p value 0.02	Yes - p value 0.0
WDAY	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
XEL	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0
ZS	Yes - p value 0.0	Yes - p value 0.0	Yes - p value 0.0

Figure 10: Stationarity Co-integration Test Results

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ABNB	0.719498	0.528065	0.320793	0.304056	0.489911	0.556802	0.817015	0.463864	0.157771	0.306964	0.414709	0.313209	0.984207	0.853235
ADBE	0.540307	0.835848	0.434041	0.766896	0.411330	0.024454	0.380866	0.890985	0.270236	0.235004	0.620457	0.622516	0.889459	0.846270
ADI	0.421497	0.810314	0.801588	0.342022	0.539857	0.802398	0.727279	0.010134	0.048431	0.893523	0.107625	0.489730	0.765676	0.329058
ADP	0.163492	0.589522	0.846160	0.469225	0.501388	0.958509	0.314213	0.911787	0.765644	0.633627	0.810423	0.108620	0.300321	0.986046
ADSK	0.979080	0.663658	0.723082	0.809696	0.374085	0.191935	0.683547	0.845142	0.076722	0.361006	0.245236	0.405117	0.243655	0.363946
AEP	0.579713	0.864165	0.488367	0.388275	0.056521	0.364422	0.454359	0.060657	0.345737	0.993195	0.462232	0.983063	0.053023	0.046015
ALGN	0.096146	0.612740	0.001876	0.110059	0.861079	0.770439	0.171364	0.069537	0.424410	0.512752	0.615711	0.804206	0.754793	0.885757
AMAT	0.542902	0.075662	0.163614	0.953313	0.528707	0.235818	0.818244	0.860731	0.291846	0.202857	0.195395	0.651440	0.963623	0.837808
AMGN	0.333614	0.714524	0.914448	0.702825	0.602644	0.693084	0.839888	0.048726	0.142144	0.554236	0.076894	0.798073	0.643447	0.431443
ANSS	0.291222	0.315005	0.990713	0.216767	0.877856	0.182948	0.981345	0.319520	0.899160	0.568517	0.171550	0.151212	0.783183	0.626001
ASML	0.517577	0.657538	0.491381	0.616493	0.803915	0.420547	0.401720	0.875389	0.934917	0.507415	0.948797	0.381527	0.372535	0.075738
ATVI	0.934038	0.242444	0.023067	0.091878	0.739574	0.112208	0.139652	0.846933	0.358721	0.810353	0.347069	0.466401	0.257401	0.119381
AVGO	0.043347	0.556209	0.993140	0.775175	0.964532	0.229409	0.196371	0.234375	0.971764	0.809268	0.863908	0.968695	0.336328	0.172411
BIDU	0.149365	0.587971	0.664120	0.824100	0.111670	0.157101	0.514831	0.174809	0.069101	0.327103	0.308348	0.894908	0.810744	0.018487
BIIB	0.790427	0.324661	0.922302	0.633291	0.970501	0.935975	0.623960	0.031423	0.070939	0.713255	0.265146	0.773510	0.111824	0.118941
BKNG	0.511109	0.540325	0.631836	0.970953	0.512563	0.409544	0.223464	0.053656	0.132253	0.157842	0.397868	0.500303	0.312547	0.533801
CDNS	0.772937	0.797308	0.098427	0.135927	0.558565	0.665469	0.082101	0.060694	0.119594	0.021041	0.336591	0.087062	0.081612	0.288256
CHTR	0.978425	0.505141	0.830193	0.348954	0.198542	0.844327	0.616753	0.803022	0.938758	0.359414	0.494674	0.212410	0.939035	0.062697
CMCSA	0.299040	0.479239	0.608994	0.679231	0.138350	0.845481	0.249227	0.739488	0.982842	0.625560	0.464395	0.452005	0.385477	0.054951
CME	0.122332	0.950009	0.814963	0.405297	0.429907	0.936315	0.367806	0.227741	0.474351	0.585548	0.065389	0.462139	0.809149	0.193164
COIN	0.171916	0.907998	0.061725	0.782147	0.838729	0.931077	0.244036	0.012483	0.008267	0.590610	0.503017	0.151609	1.000000	0.000000
COST	0.769718	0.179457	0.000403	0.142193	0.240057	0.662074	0.211207	0.606910	0.618746	0.593929	0.343320	0.035693	0.532151	0.575246
CPRT	0.071516	0.284412	0.941475	0.645019	0.077565	0.087180	0.551373	0.505361	0.107181	0.632209	0.867750	0.486306	0.584996	0.404628
CRWD	0.378921	0.467975	0.904541	0.160511	0.258405	0.144823	0.493943	0.559371	0.486959	0.950099	0.405406	0.351332	0.339186	0.718834
CSCO	0.415911	0.817331	0.829330	0.967214	0.237183	0.025110	0.248020	0.418375	0.279645	0.295593	0.073241	0.299854	0.437220	0.049244
CSX	0.003146	0.912598	0.105333	0.316902	0.488978	0.125795	0.341746	0.603558	0.986674	0.672990	0.713845	0.768608	0.143975	0.059977
CTAS	0.952120	0.273346	0.079140	0.790441	0.060115	0.794528	0.017238	0.716803	0.061959	0.088227	0.775754	0.052093	0.230348	0.668525
CTSH	0.563208	0.345891	0.499067	0.912467	0.891037	0.914296	0.300972	0.916143	0.310703	0.711016	0.437901	0.924384	0.187646	0.842104
DDOG	0.909305	0.372907	0.235364	0.119692	0.877632	0.402556	0.930097	0.780477	0.798518	0.911904	0.760915	0.648487	0.442406	0.756199
DLTR	0.698020	0.426864	0.935158	0.601128	0.943899	0.731984	0.585733	0.032235	0.038613	0.749596	0.698012	0.771673	0.768883	0.613840
DOCU	0.484784	0.597840	0.242344	0.128304	0.991160	0.068956	0.658414	0.257106	0.916057	0.458326	0.656785	0.952081	0.516088	0.700869
DXCM	0.433167	0.783471	0.782797	0.872262	0.943703	0.456824	0.810927	0.789070	0.952137	0.889971	0.460476	0.679411	0.723505	0.513723
EA	0.723972	0.863166	0.549844	0.951365	0.461588	0.825970	0.650172	0.630614	0.256084	0.277536	0.596505	0.322425	0.548554	0.748769
EBAY	0.220817	0.639247	0.119462	0.270935	0.116022	0.087067	0.858403	0.332032	0.595402	0.429378	0.349117	0.713986	0.607580	0.664193
ETSY	0.436888	0.212827	0.189830	0.644148	0.852115	0.243876	0.483695	0.265210	0.073856	0.702962	0.498113	0.292382	0.001790	0.407989
EXC	0.343310	0.005615	0.539378	0.174219	0.114734	0.556589	0.378241	0.282079	0.619206	0.985789	0.901059	0.581270	0.363598	0.770646
EXPE	0.614189	0.913601	0.847183	0.648085	0.240766	0.808231	0.682944	0.111045	0.002909	0.093791	0.260305	0.051531	0.913217	0.972281
FAST	0.738103	0.348905	0.924773	0.228477	0.074722	0.106336	0.389386	0.561495	0.576471	0.983641	0.188523	0.425597	0.851197	0.550218
FISV	0.008021	0.084696	0.332885	0.527312	0.714128	0.625848	0.340500	0.544946	0.448215	0.753351	0.869982	0.607538	0.211720	0.040570
FTNT	0.856140	0.311362	0.756431	0.381553	0.512653	0.491397	0.876860	0.480983	0.312603	0.187699	0.422797	0.737773	0.587358	0.538334
GILD	0.256224	0.762135	0.588061	0.210067	0.911150	0.522294	0.494311	0.365211	0.820192	0.890295	0.342272	0.126178	0.933706	0.268557
HON	0.012934	0.639086	0.374925	0.746180	0.953529	0.555674	0.064098	0.110278	0.737068	0.873033	0.819919	0.199437	0.951247	0.473322
IDXX	0.749398	0.814928	0.980221	0.808659	0.481706	0.979948	0.890146	0.714350	0.391050	0.822167	0.349254	0.751951	0.926049	0.363858
ILMN	0.980826	0.405892	0.416207	0.209718	0.018144	0.391168	0.671016	0.329122	0.515203	0.794873	0.162332	0.655433	0.898499	0.701812
INTU	0.817398	0.490527	0.092657	0.693141	0.963773	0.406274	0.870245	0.767003	0.759445	0.850079	0.368012	0.948215	0.928296	0.861039
ISRG	0.284430	0.020201	0.043459	0.598723	0.009859	0.002681	0.081031	0.808717	0.093446	0.134599	0.270314	0.782073	0.598511	0.156908
JD	0.625301	0.622712	0.387506	0.332018	0.962390	0.242389	0.084035	0.398985	0.313311	0.204223	0.467708	0.685410	0.353716	0.875528

Figure 11: Correlation Test Results - 1

KDP	0.857324	0.851771	0.387633	0.089019	0.009448	0.587402	0.462805	0.308743	0.460376	0.016108	0.063729	0.129116	0.230946	0.602656
KHC	0.010006	0.948091	0.538047	0.322580	0.441079	0.554647	0.088569	0.934070	0.500591	0.584044	0.500835	0.525978	0.246612	0.893398
KLAC	0.974382	0.953066	0.451750	0.012684	0.032998	0.792483	0.539991	0.717600	0.879905	0.457210	0.306635	0.512073	0.590904	0.856487
LCID	0.084044	0.905319	0.351675	0.038767	0.327171	0.404970	0.860120	0.355336	0.636698	0.479509	0.660172	0.412727	0.141484	0.887377
LPL	0.954139	0.403965	0.066780	0.334307	0.554534	0.475308	0.710386	0.739114	0.991583	0.247034	0.879227	0.887484	0.788272	0.561693
LRCX	0.824028	0.629932	0.182578	0.134165	0.021190	0.034143	0.024996	0.376124	0.968758	0.229480	0.599514	0.990582	0.406363	0.497262
LULU	0.058203	0.445170	0.850825	0.398086	0.706837	0.715039	0.904822	0.375173	0.059529	0.039141	0.549346	0.979915	0.829608	0.375599
MAR	0.004807	0.005481	0.559525	0.893019	0.562580	0.725930	0.120517	0.567446	0.729926	0.373350	0.253403	0.183243	0.830931	0.458215
MCHP	0.556676	0.103463	0.716956	0.020375	0.198460	0.641888	0.204707	0.788184	0.861353	0.817397	0.735267	0.683663	0.503593	0.605167
MDLZ	0.606096	0.944579	0.477273	0.997928	0.597863	0.212936	0.780710	0.986002	0.038052	0.702057	0.005824	0.661515	0.023212	0.201641
MELI	0.690009	0.660233	0.084851	0.406841	0.283231	0.383558	0.289671	0.635990	0.409467	0.877913	0.169960	0.305358	0.568926	0.548795
MNST	0.188729	0.493955	0.906035	0.481834	0.037906	0.143339	0.511037	0.950777	0.714067	0.770444	0.110715	0.196640	0.013864	0.369718
MRNA	0.576029	0.151736	0.954610	0.552027	0.374846	0.481033	0.744109	0.203728	0.995523	0.321460	0.735937	0.603988	0.710080	0.998398
MRVL	0.100664	0.093378	0.151803	0.946610	0.655003	0.771492	0.439061	0.179866	0.619048	0.936884	0.090252	0.051382	0.954071	0.261152
MTCH	0.473006	0.256005	0.313354	0.438023	0.759932	0.595486	0.778968	0.844451	0.128786	0.025540	0.153093	0.308780	0.389363	0.244904
MU	0.226114	0.157454	0.307439	0.114193	0.473895	0.505472	0.923421	0.204755	0.037714	0.978570	0.390464	0.236271	0.692241	0.350500
NTES	0.691433	0.252391	0.943070	0.346884	0.657071	0.645875	0.823883	0.406087	0.374533	0.523374	0.855951	0.876978	0.328602	0.109690
NTRS	0.075701	0.015022	0.850363	0.197022	0.028816	0.657882	0.589143	0.812553	0.661402	0.731782	0.824547	0.857664	0.730260	0.965621
NXPI	0.239406	0.614534	0.930537	0.813143	0.262084	0.716018	0.052110	0.008634	0.687231	0.288507	0.068047	0.121823	0.303316	0.342422
ODFL	0.260485	0.561516	0.402016	0.977683	0.841508	0.815178	0.370491	0.144014	0.704519	0.673599	0.393224	0.596701	0.427508	0.445336
OKTA	0.930036	0.084101	0.902981	0.556796	0.584382	0.321209	0.259923	0.778027	0.684134	0.740230	0.454556	0.071579	0.502944	0.036594
ORLY	0.345481	0.522318	0.789955	0.973169	0.062468	0.959077	0.758892	0.155262	0.126436	0.394253	0.104322	0.272422	0.150735	0.245901
PANW	0.170286	0.154054	0.312912	0.348423	0.354812	0.353235	0.830834	0.580973	0.825704	0.437843	0.973247	0.661903	0.962232	0.559235
PAYX	0.100329	0.656667	0.340065	0.929644	0.125368	0.275280	0.234453	0.937094	0.385353	0.029175	0.042058	0.436565	0.432578	0.330807
PCAR	0.608658	0.196892	0.024328	0.149380	0.397352	0.878926	0.815642	0.321033	0.735200	0.331163	0.011684	0.979981	0.347924	0.467055
PDD	0.053756	0.147439	0.324789	0.532972	0.763787	0.927606	0.844262	0.490528	0.661965	0.446638	0.337203	0.209873	0.891136	0.562335
PEP	0.636209	0.093142	0.957734	0.842955	0.830216	0.164536	0.965995	0.933523	0.654285	0.376390	0.572587	0.343538	0.384797	0.456907
PTON	0.990456	0.783302	0.501365	0.107943	0.090393	0.801191	0.902524	0.878809	0.596350	0.382877	0.587842	0.574474	0.858373	0.677052
PYPL	0.609885	0.141975	0.618910	0.069683	0.455619	0.051548	0.484145	0.764158	0.752874	0.741218	0.675479	0.340264	0.251763	0.280030
QCOM	0.443315	0.927516	0.793371	0.073044	0.011222	0.693305	0.528822	0.651588	0.418427	0.630220	0.443638	0.077338	0.334192	0.441468
REGN	0.364892	0.929528	0.506536	0.482400	0.731929	0.310056	0.467765	0.833020	0.791739	0.737951	0.769679	0.905916	0.634333	0.278959
RIVN	0.314823	0.137924	0.349982	0.482186	0.734977	0.530963	0.602385	0.945129	0.994767	0.841893	0.645499	0.531397	0.453645	0.227432
ROST	0.227479	0.265576	0.649144	0.513616	0.822069	0.138578	0.030447	0.680467	0.348977	0.503696	0.620468	0.081528	0.624032	0.468393
SGEN	0.289987	0.715844	0.365738	0.236228	0.012681	0.000806	0.342366	0.167918	0.970745	0.321879	0.282138	0.523313	0.734098	0.367014
SIRI	0.398828	0.548380	0.661203	0.557190	0.562646	0.447743	0.938789	0.915266	0.955507	0.841189	0.809212	0.500473	0.168892	0.473508
SNPS	0.055494	0.988800	0.214554	0.019789	0.325555	0.494635	0.006163	0.794209	0.001729	0.078432	0.349525	0.115881	0.012146	0.238309
SPLK	0.188280	0.536599	0.256367	0.108944	0.688059	0.431632	0.407694	0.400714	0.888702	0.547676	0.525717	0.149239	0.087915	0.580152
SWKS	0.665547	0.621012	0.812961	0.275539	0.293345	0.304983	0.179979	0.249240	0.323402	0.735464	0.232934	0.125410	0.910675	0.134131
TEAM	0.010945	0.889782	0.185293	0.312980	0.653791	0.895570	0.705053	0.267659	0.021692	0.560586	0.174293	0.876919	0.717335	0.726673
TMUS	0.911109	0.463536	0.194764	0.433614	0.443108	0.184602	0.203144	0.258530	0.151411	0.633248	0.767683	0.267068	0.030619	0.490962
TXN	0.551324	0.847506	0.940166	0.223905	0.416446	0.623477	0.298731	0.998239	0.109982	0.591649	0.120139	0.040914	0.334657	0.134423
VALE	0.055236	0.008091	0.585828	0.280430	0.732085	0.311836	0.957026	0.060883	0.057050	0.628380	0.599667	0.025307	0.132896	0.253274
VRSK	0.886868	0.370959	0.524342	0.958993	0.361485	0.396176	0.790183	0.445529	0.310600	0.394974	0.672817	0.277238	0.574305	0.307588
VRSN	0.197012	0.149034	0.844357	0.873354	0.138385	0.298244	0.915618	0.578559	0.424567	0.574812	0.969417	0.259391	0.991841	0.746444
VRTX	0.517332	0.448923	0.033932	0.283559	0.728726	0.398919	0.446032	0.979295	0.721080	0.903946	0.599858	0.211445	0.004401	0.301707
WBA	0.396400	0.659579	0.934793	0.542750	0.672489	0.585399	0.850409	0.199677	0.787537	0.454176	0.479333	0.072033	0.829476	0.305776
WDAY	0.811193	0.220404	0.324410	0.592452	0.901942	0.290719	0.278473	0.297621	0.437659	0.240636	0.249419	0.564714	0.463562	0.975990
XEL	0.587339	0.512884	0.282770	0.276778	0.336856	0.300292	0.888925	0.415959	0.710630	0.169270	0.108505	0.389353	0.575542	0.781743
ZM	0.831148	0.948138	0.115982	0.183214	0.651166	0.084049	0.035018	0.235047	0.768498	0.705947	0.292986	0.654527	0.527241	0.261436
ZS	0.052870	0.449040	0.740507	0.901999	0.587700	0.263241	0.982235	0.870862	0.798827	0.814355	0.254403	0.412954	0.892988	0.222500

Figure 12: Correlation Test Results - 2

	Min_AIC_score	Min_RMSE_score	Orders	Min_AIC_Order	Min_RMSE_Order
ADP	-261.352204	0.035203	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 3)
ALGN	183.659012	0.065898	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 1)
AVGO	101.814215	0.262085	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 1)
BIDU	-9.418418	0.033286	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 1)
CHTR	162.220825	0.069106	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 3)	(1, 0)
CME	158.210887	0.063134	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(2, 2)
COIN	-66.830254	0.034530	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(4, 0)	(1, 2)
COST	85.745037	0.132291	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 2)
CPRT	-59.898217	0.033368	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(1, 1)	(0, 1)
CSX	-131.761218	0.034500	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 0)
EBAY	-126.212507	0.031199	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(4, 0)	(1, 4)
EXC	-133.825501	0.030738	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(2, 3)
FISV	-103.551206	0.036509	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(2, 4)
HON	59.798480	0.059528	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 1)
IDXX	41.228776	0.040457	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 1)
ILMN	-226.084991	0.033500	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(4, 0)
ISRG	-254.168840	0.032890	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 1)
KDP	-129.887768	0.031867	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(4, 0)	(0, 1)
KHC	-247.457949	0.033948	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(1, 4)	(0, 4)
LCID	26.141265	0.041075	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 0)
LPL	-12.544493	0.039529	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(4, 4)
LRCX	627.801655	0.465278	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(2, 2)
LULU	-76.020872	0.031430	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 3)
MAR	-78.216055	0.030303	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 2)	(0, 2)
MRNA	8.451527	0.043665	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(4, 1)
MRVL	-141.597794	0.032082	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 3)
MU	-92.810026	0.033500	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 3)
NTRS	-355.456892	0.035285	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 2)	(1, 0)
PAYX	-82.150974	0.030438	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 2)
PCAR	-128.096568	0.031397	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(4, 0)	(2, 2)
PDD	-86.819912	0.033116	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 4)
PEP	77.353055	0.035530	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 3)
QCOM	-22.478693	0.042715	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(2, 1)
SNPS	620.998847	0.583134	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 1)
TEAM	2.510161	0.037875	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(0, 2)
TMUS	-51.846648	0.031923	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 0)
TXN	-117.120359	0.034514	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(3, 2)
VALE	-61.964714	0.033615	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 0)
VRSN	-79.885546	0.034079	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(3, 0)	(2, 2)
WDAY	196.821017	0.071987	[(0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1...	(0, 1)	(1, 1)
XEL	0.000000	0.000000	0.0	0.0	0.0
ZS	0.000000	0.000000	0.0	0.0	0.0

Figure 13: Minimum AIC Score RMSE for different order VARMA Models

	Min_AIC_score	RMSE_score	AIC Lag	RMSE Lag
ADP	-9.885052	0.028977	3.0	3.0
ALGN	-8.360177	0.069681	3.0	1.0
AVGO	-137.229281	0.021183	7.0	1.0
BIDU	-6.859644	0.045288	2.0	5.0
CHTR	-9.505217	0.031451	2.0	1.0
CME	-10.276253	0.024488	4.0	8.0
COIN	-139.779667	0.033310	6.0	1.0
COST	-137.774162	0.021047	8.0	1.0
CPRT	-9.084406	0.025592	2.0	5.0
CSX	-9.290756	0.021705	3.0	1.0
EBAY	-9.689387	0.016041	4.0	1.0
EXC	-10.332874	0.029036	2.0	2.0
FISV	-9.888364	0.022823	5.0	3.0
HON	-10.429507	0.013654	6.0	2.0
IDXX	-9.089735	0.037540	3.0	6.0
ILMN	-8.931540	0.032913	7.0	2.0
ISRG	-9.326059	0.050542	2.0	3.0
KDP	-9.987076	0.016078	4.0	7.0
KHC	-10.659917	0.014237	8.0	3.0
LCID	-7.651556	0.047148	8.0	3.0
LPL	-133.260253	0.075197	8.0	6.0
LRCX	-9.032308	0.026344	3.0	4.0
LULU	-11.797572	0.026258	8.0	4.0
MAR	-9.610832	0.029959	2.0	7.0
MRNA	-139.728096	0.037485	7.0	1.0
MRVL	-8.926900	0.032416	3.0	7.0
MU	-139.754609	0.028865	8.0	2.0
NTRS	-9.456801	0.015345	5.0	4.0
PAYX	-9.682219	0.023599	1.0	5.0
PCAR	-9.618096	0.031036	8.0	1.0
PDD	-6.765583	0.083632	2.0	1.0
PEP	-13.552441	0.006896	8.0	5.0
QCOM	-138.140551	0.033828	7.0	3.0
SNPS	-8.884936	0.030764	7.0	2.0
TEAM	-7.759580	0.058989	1.0	6.0
TMUS	-10.758943	0.017799	8.0	3.0
TXN	-10.224901	0.019867	4.0	6.0
VALE	-8.313539	0.036792	7.0	6.0
VRSN	-137.334722	0.039300	8.0	1.0
WDAY	-8.357708	0.039587	2.0	1.0
XEL	-10.237605	0.014345	2.0	1.0
ZS	-7.926585	0.033565	3.0	2.0

Figure 14: Minimum AIC Score RMSE for different order VAR Models

	Min_RMSE_score VARMA	Min_RMSE_score VAR
ADP	0.035203	0.028977
ALGN	0.065898	0.069681
AVGO	0.262085	0.021183
BIDU	0.033286	0.045288
CHTR	0.069106	0.031451
CME	0.063134	0.024488
COIN	0.034530	0.033310
COST	0.132291	0.021047
CPRT	0.033368	0.025592
CSX	0.034500	0.021705
EBAY	0.031199	0.016041
EXC	0.030738	0.029036
FISV	0.036509	0.022823
HON	0.059528	0.013654
IDXX	0.040457	0.037540
ILMN	0.033500	0.032913
ISRG	0.032890	0.050542
KDP	0.031867	0.016078
KHC	0.033948	0.014237
LCID	0.041075	0.047148
LPL	0.039529	0.075197
LRCX	0.465278	0.026344
LULU	0.031430	0.026258
MAR	0.030303	0.029959
MRNA	0.043665	0.037485
MRVL	0.032082	0.032416
MU	0.033500	0.028865
NTRS	0.035285	0.015345
PAYX	0.030438	0.023599
PCAR	0.031397	0.031036
PDD	0.033116	0.083632
PEP	0.035530	0.006896
QCOM	0.042715	0.033828
SNPS	0.583134	0.030764
TEAM	0.037875	0.058989
TMUS	0.031923	0.017799
TXN	0.034514	0.019867
VALE	0.033615	0.036792
VRSN	0.034079	0.039300
WDAY	0.071987	0.039587
XEL	0.000000	0.014345
ZS	0.000000	0.033565

Figure 15: Comparing VAR VARMA Models

Statespace Model Results

Dep. Variable:	['diff_title_score', 'diff_stock_price']	No. Observations:	63
Model:	VMA(1) + intercept	Log Likelihood	37.477
		AIC	-56.955
Date:	Thu, 12 May 2022	BIC	-37.666
Time:	08:48:54	HQIC	-49.368
Sample:	0 - 63		
Covariance Type:	opg		

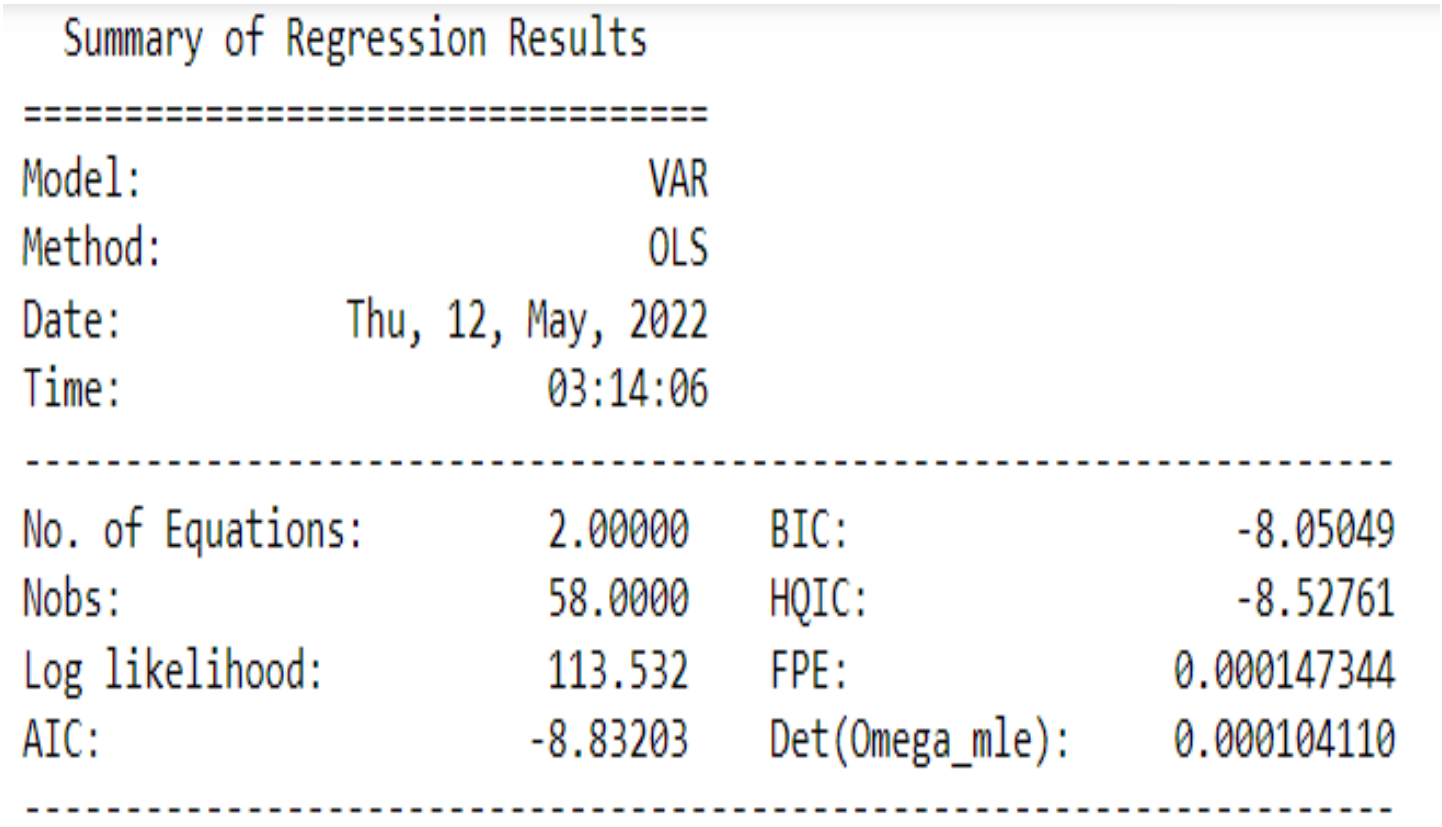
Ljung-Box (L1) (Q):	0.02, 20.32	Jarque-Bera (JB):	1.79, 759.60
Prob(Q):	0.89, 0.00	Prob(JB):	0.41, 0.00
Heteroskedasticity (H):	0.74, 0.11	Skew:	-0.08, -3.45
Prob(H) (two-sided):	0.49, 0.00	Kurtosis:	2.19, 18.55

Figure 16: VARMA Model for CPRT stock-1

Results for equation diff_stock_price

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.0001	0.006	-0.017	0.987	-0.012	0.012
L1.e(diff_title_score)	0.0722	0.058	1.252	0.210	-0.041	0.185
L1.e(diff_stock_price)	-0.7800	0.197	-3.950	0.000	-1.167	-0.393

Figure 17: VARMA Model for CPRT stock-2



```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 12, May, 2022
Time:                      03:14:06
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   -8.05049
Nobs:                     58.0000    HQIC:                  -8.52761
Log likelihood:           113.532    FPE:                0.000147344
AIC:                     -8.83203    Det(Omega_mle):     0.000104110
--------------------------------------------------------------------
```

Figure 18: VAR Model for CPRT stock-1

Results for equation diff_stock_price

	coefficient	std. error	t-stat	prob
const	0.001657	0.004209	0.394	0.694
L1.diff_title_score	0.021620	0.013737	1.574	0.116
L1.diff_stock_price	0.001884	0.143403	0.013	0.990
L2.diff_title_score	0.004395	0.018417	0.239	0.811
L2.diff_stock_price	-0.141228	0.151628	-0.931	0.352
L3.diff_title_score	-0.007157	0.020406	-0.351	0.726
L3.diff_stock_price	-0.005280	0.165283	-0.032	0.975
L4.diff_title_score	-0.018333	0.018944	-0.968	0.333
L4.diff_stock_price	0.051442	0.173741	0.296	0.767
L5.diff_title_score	-0.024565	0.014006	-1.754	0.079
L5.diff_stock_price	-0.015048	0.045726	-0.329	0.742

Figure 19: VAR Model for CPRT stock-2

References

Batra, R. and Daudpota, S. M. (2018), Integrating stocktwits with sentiment analysis for better prediction of stock price movement, *in* '2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)', IEEE, pp. 1–5.

Boh, D. (2022), 'Sentiment analysis of stocks from financial news using python'.
URL: *https://medium.datadriveninvestor.com/sentiment-analysis-of-stocks-from-financial-news-using-python-82ebdcefb638*

Chen, C.-C., Huang, H.-H. and Chen, H.-H. (2018), Ntusd-fin: A market sentiment dictionary for financial social media data applications.

Chen, L. (2021), Fixed effect regression — simply explained.
URL: *https://towardsdatascience.com/fixed-effect-regression-simply-explained-ab690bd885cf*

Cointegration tests on time series (n.d.).
URL: *https://medium.com/bluekiri/cointegration-tests-on-time-series-88702ea9c492*

Fataliyev, K., Chivukula, A., Prasad, M. and Liu, W. (2021), 'Stock market analysis with text data: A review', *arXiv:2106.12985 [cs, q-fin]* . arXiv: 2106.12985.
URL: *http://arxiv.org/abs/2106.12985*

How to Perform a Granger-Causality Test in Python (n.d.).
URL: *https://www.statology.org/granger-causality-test-in-python/*

Hutto, C. and Gilbert, E. (2014), 'Vader: A parsimonious rule-based model for sentiment analysis of social media text', *Proceedings of the International AAAI Conference on Web and Social Media* **8**(1), 216–225.
URL: *https://ojs.aaai.org/index.php/ICWSM/article/view/14550*

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, 5 edn, Springer Berlin Heidelberg, Berlin, Heidelberg.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D. C. (2019), Stock price prediction using news sentiment analysis, *in* '2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)', pp. 205–208.

PYTHON FOR DATA SCIENCE (n.d.).

URL: *https://www.pythonfordatascience.org/mixed-effects-regression-python/*

Uhr, P., Zenkert, J. and Fathi, M. (2014), Sentiment analysis in financial markets a framework to utilize the human ability of word association for analyzing stock market news reports, *in* '2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)', pp. 912–917.