The Impact of News Sentiment on MSFT

Ansam Zedan, Daniel Wullschleger

2023-12-18

Introduction

In recent years Microsoft has pushed and achieved great progress in a diverse portfolio of technologies (Microsoft, 2022) including also partnerships with AI research centers such as OpenAI. Alongside this development, its stock price has since reached new all-time highs and persists to be a major player in tech investments.

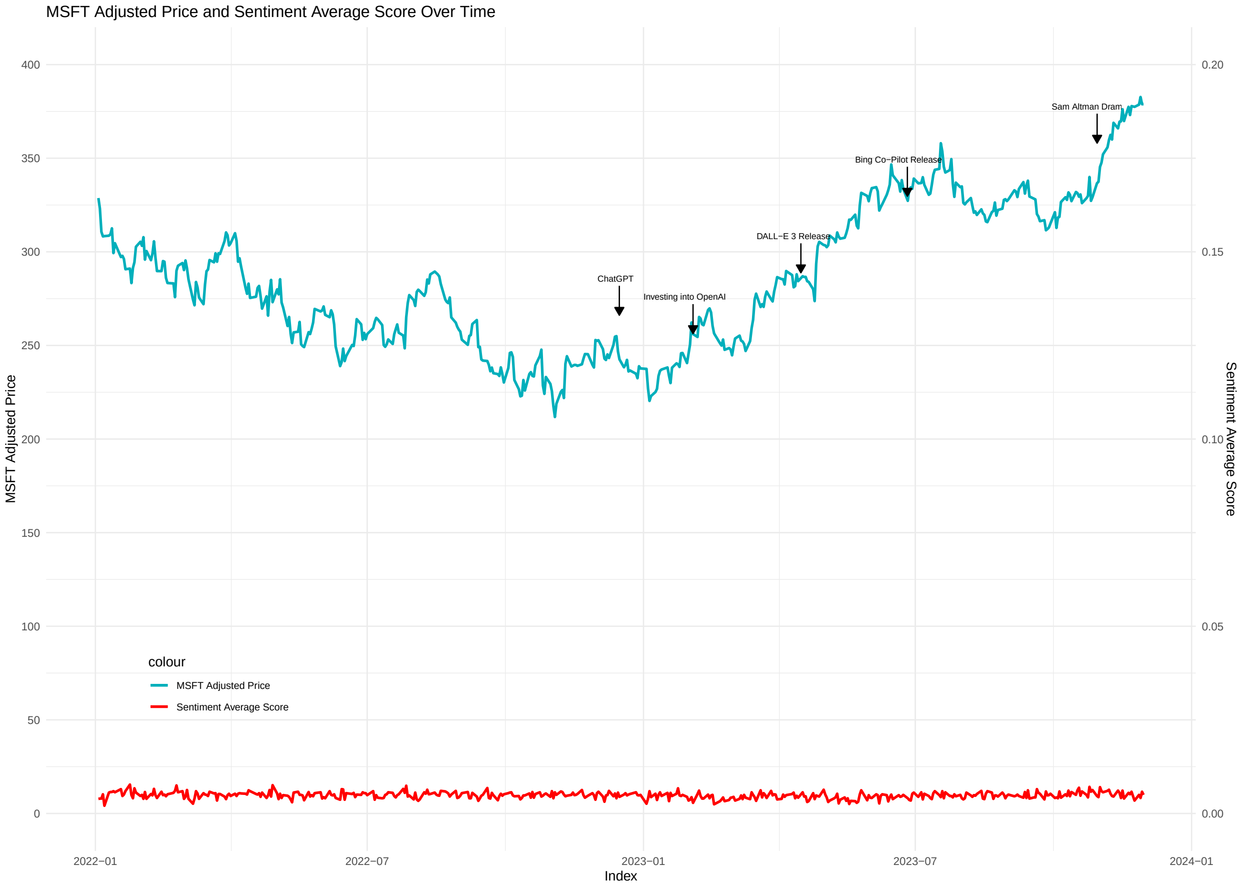
Inspired by this evolution, this paper provides an analysis and forecast of the stock price of Microsoft based on news sentiment data from subreddits such as r/worldnews, r/investing and r/stocks using a vector autoregressive (VAR) model.

Vector Autoregressive (VAR) processes are popular in economics for their flexibility and simplicity as models for multivariate time series data, forming a vector between variables that affect each other (Suharsono et al. 2017; Aptech 2021). The relationship between the stock prices and the news events will be analysed using a Granger causality testing, similarly as Bhowmik et al. (2022) proceeded when comparing multiple markets.

Hereby MSFT price data was downloaded through the Yahoo API in R and Reddit posts were scraped using Jupyter Notebooks. The python libraries nltk, textblob and wordlcoud were of support in the creation of the sentiment scores.

Analysis:

First, we import the data and convert it to a time series so that we may plot it. After importing and converting our dataset into a workable format, we observed noticeable noise in the sentiment_average_score data. To address this, we applied a LOESS smoothing technique. This method allows us to smooth out short-term fluctuations and highlight longer-term trends in sentiment data, crucial for correlating with stock price movements. The choice of the smoothing parameter (span) was iteratively refined to balance detail and smoothness

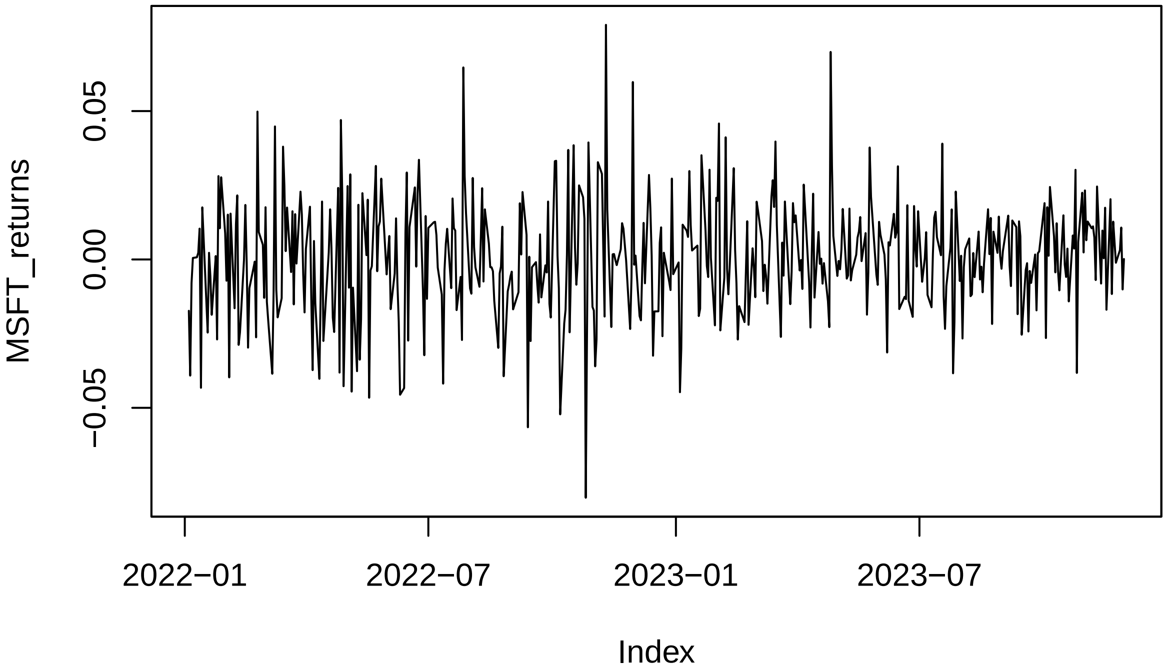


We see a clear trend, which is an indicator for non-stationarity, so we do an Augmented Dickey-Fuller (ADF) Test to see whether the data is in fact non-stationary.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  MSFT_ts
## Dickey-Fuller = -1.6998, Lag order = 7, p-value = 0.7051
## alternative hypothesis: stationary
```

Achieving Stationarity

We see that based on the p-value of 0.7051 the ADF Test suggests that the price data is non-stationary. So we a apply a log-transform with subsequent differencing to calculate continuous returns on MSFT. These should behave stationary.



```
##
##  Augmented Dickey-Fuller Test
##
## data:  MSFT_returns
## Dickey-Fuller = -8.874, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Creating a VAR Model using News sentiment

Now, that we deal with stationary data we can begin with our analysis.

We divide the dataset into a Train/Test Split, whereas November and December 2023 will serve as a test set for later predicitions.

First we create a time series of news sentiment data. Combined with the price returns, we can create a Vector Autoregression model. This will be our basis to check whether News are a causal driver for MSFT price returns.

We then do a coefficient test, to get more information on the statistical significance of the coefficients of our VAR Model.

```
##
## t test of coefficients:
##
##                                       Estimate  Std. Error t value  Pr(>|t|)
## MSFT_Returns:(Intercept)            2.4446e-02  1.3766e-02  1.7758 0.0765378 .
## MSFT_Returns:MSFT_Returns.l1      -4.8719e-03  5.0449e-02 -0.0966 0.9231164    
## MSFT_Returns:News_Sentiment.l1     8.5003e-04  6.0794e-04  1.3982 0.1628392    
## MSFT_Returns:MSFT_Returns.l2      -1.1504e-01  5.0413e-02 -2.2819 0.0230310 *  
## MSFT_Returns:News_Sentiment.l2    -5.9825e-04  5.9998e-04 -0.9971 0.3193248    
## MSFT_Returns:MSFT_Returns.l3      -7.4059e-02  5.0401e-02 -1.4694 0.1425266    
## MSFT_Returns:News_Sentiment.l3     5.8379e-04  5.9344e-04  0.9837 0.3258534    
## MSFT_Returns:MSFT_Returns.l4       5.4752e-03  5.0682e-02  0.1080 0.9140274    
## MSFT_Returns:News_Sentiment.l4     4.9556e-06  5.8954e-04  0.0084 0.9932974    
## MSFT_Returns:MSFT_Returns.l5       4.8317e-02  5.0316e-02  0.9603 0.3375099    
## MSFT_Returns:News_Sentiment.l5    -1.2643e-03  5.9492e-04 -2.1252 0.0341986 *  
## MSFT_Returns:MSFT_Returns.l6      -6.6429e-02  5.0171e-02 -1.3240 0.1862628    
## MSFT_Returns:News_Sentiment.l6    -6.4916e-04  5.9547e-04 -1.0902 0.2763161    
## MSFT_Returns:MSFT_Returns.l7       1.6465e-02  5.0306e-02  0.3273 0.7436182    
## MSFT_Returns:News_Sentiment.l7     3.7340e-04  5.9538e-04  0.6272 0.5309137    
## MSFT_Returns:MSFT_Returns.l8      -7.6278e-02  4.9737e-02 -1.5336 0.1259307    
## MSFT_Returns:News_Sentiment.l8    -1.6030e-03  5.9601e-04 -2.6896 0.0074607 ** 
## MSFT_Returns:MSFT_Returns.l9       6.3871e-02  4.9889e-02  1.2803 0.2012149    
## MSFT_Returns:News_Sentiment.l9    -2.9930e-04  6.0202e-04 -0.4972 0.6193593    
## MSFT_Returns:MSFT_Returns.l10     -5.4269e-02  4.9897e-02 -1.0876 0.2774292    
## MSFT_Returns:News_Sentiment.l10    4.4956e-04  5.9844e-04  0.7512 0.4529733    
## MSFT_Returns:MSFT_Returns.l11     -1.0655e-01  4.9370e-02 -2.1582 0.0315230 *  
## MSFT_Returns:News_Sentiment.l11    8.0541e-04  6.0544e-04  1.3303 0.1841988    
## MSFT_Returns:MSFT_Returns.l12     -7.7592e-02  4.9646e-02 -1.5629 0.1188855    
## MSFT_Returns:News_Sentiment.l12    1.0627e-03  5.9088e-04  1.7986 0.0728559 .  
## MSFT_Returns:MSFT_Returns.l13      6.9952e-02  4.9757e-02  1.4059 0.1605568    
## MSFT_Returns:News_Sentiment.l13   -4.5320e-04  5.9895e-04 -0.7567 0.4497065    
## MSFT_Returns:MSFT_Returns.l14     -1.7271e-02  4.9835e-02 -0.3466 0.7291084    
## MSFT_Returns:News_Sentiment.l14   -2.6401e-04  5.9965e-04 -0.4403 0.6599836    
## MSFT_Returns:MSFT_Returns.l15     -3.9022e-02  4.9976e-02 -0.7808 0.4353846    
## MSFT_Returns:News_Sentiment.l15   -1.5122e-03  5.9566e-04 -2.5387 0.0115146 *  
## News_Sentiment:(Intercept)         4.8040e+00  1.1475e+00  4.1864 3.504e-05 ***
## News_Sentiment:MSFT_Returns.l1    -1.0947e+00  4.2054e+00 -0.2603 0.7947655    
## News_Sentiment:News_Sentiment.l1   9.5363e-02  5.0677e-02  1.8818 0.0606076 .  
## News_Sentiment:MSFT_Returns.l2    -3.0413e+00  4.2024e+00 -0.7237 0.4696791    
## News_Sentiment:News_Sentiment.l2   1.2257e-02  5.0013e-02  0.2451 0.8065313    
## News_Sentiment:MSFT_Returns.l3    -9.5139e+00  4.2014e+00 -2.2645 0.0240912 *  
## News_Sentiment:News_Sentiment.l3   4.8679e-02  4.9468e-02  0.9840 0.3257055    
## News_Sentiment:MSFT_Returns.l4    -1.6041e+00  4.2248e+00 -0.3797 0.7043893    
## News_Sentiment:News_Sentiment.l4   1.2787e-01  4.9143e-02  2.6019 0.0096220 ** 
## News_Sentiment:MSFT_Returns.l5    -2.0665e+00  4.1943e+00 -0.4927 0.6225019    
## News_Sentiment:News_Sentiment.l5   6.8100e-02  4.9592e-02  1.3732 0.1704738    
## News_Sentiment:MSFT_Returns.l6    -6.4807e+00  4.1822e+00 -1.5496 0.1220501    
## News_Sentiment:News_Sentiment.l6  -2.3494e-02  4.9638e-02 -0.4733 0.6362606    
## News_Sentiment:MSFT_Returns.l7     2.6489e+00  4.1934e+00  0.6317 0.5279728    
## News_Sentiment:News_Sentiment.l7  -2.5793e-02  4.9630e-02 -0.5197 0.6035642    
## News_Sentiment:MSFT_Returns.l8     1.2048e+00  4.1460e+00  0.2906 0.7715227    
## News_Sentiment:News_Sentiment.l8  -6.5688e-02  4.9683e-02 -1.3221 0.1868956    
## News_Sentiment:MSFT_Returns.l9     3.7041e-01  4.1587e+00  0.0891 0.9290722    
## News_Sentiment:News_Sentiment.l9   6.2111e-02  5.0184e-02  1.2377 0.2165875    
## News_Sentiment:MSFT_Returns.l10   -5.2044e+00  4.1593e+00 -1.2512 0.2115944    
## News_Sentiment:News_Sentiment.l10  1.6680e-01  4.9885e-02  3.3437 0.0009064 ***
## News_Sentiment:MSFT_Returns.l11   -2.2978e+00  4.1154e+00 -0.5583 0.5769346    
## News_Sentiment:News_Sentiment.l11 -6.2763e-02  5.0469e-02 -1.2436 0.2143933    
## News_Sentiment:MSFT_Returns.l12   -5.2663e+00  4.1384e+00 -1.2725 0.2039373    
## News_Sentiment:News_Sentiment.l12  1.5159e-01  4.9255e-02  3.0777 0.0022331 **
## News_Sentiment:MSFT_Returns.l13   -9.5938e+00  4.1477e+00 -2.3130 0.0212388 *
## News_Sentiment:News_Sentiment.l13 -4.4498e-02  4.9928e-02 -0.8912 0.3733451
## News_Sentiment:MSFT_Returns.l14   -7.3287e+00  4.1542e+00 -1.7642 0.0784836 .
## News_Sentiment:News_Sentiment.l14 -3.6315e-02  4.9986e-02 -0.7265 0.4679638
## News_Sentiment:MSFT_Returns.l15    8.3375e+00  4.1660e+00  2.0013 0.0460476 *
## News_Sentiment:News_Sentiment.l15  2.4371e-02  4.9654e-02  0.4908 0.6238276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient test reveals some interesting insights into the dynamics between MSFT stock returns and news sentiment. In particular, several lagged values of news sentiment exhibit statistical significance, suggesting a nuanced, albeit modest, effect on stock returns. For example, both the second and fifth lags of sentiment show a negative impact, while the eighth lag shows a more pronounced negative impact. This pattern is mirrored by the eleventh and fifteenth lags, but with less statistical strength. Conversely, the impact of news sentiment is not uniform across all lags, as some show a positive relationship, although these are not statistically significant at conventional levels.

These findings suggest that while news sentiment does not consistently drive stock returns, it does have some predictive power at certain intervals. This time effect could be due to the market's lagged reaction to news sentiment or to the complex interplay between various market forces and investor psychology.

Causlaity testing

As a next step, we do a Granger causality test to confirm that our news sentiments are in fact not a causal driver for the MSFT stock price.

```
## $Granger
##
##  Granger causality H0: News_Sentiment do not Granger-cause MSFT_Returns
##
## data:  VAR object MSFT_VAR
## F-Test = 2.1049, df1 = 15, df2 = 782, p-value = 0.008237
```

he Granger causality test provided a new perspective on the relationship between news sentiment and MSFT stock returns. Contrary to our initial hypothesis, the test indicates that news sentiment may indeed have a Granger causal effect on stock returns, as indicated by a p-value of 0.008237. This result is statistically significant at conventional levels ($p < 0.05$), leading us to reconsider our previous stance. With an F-test value of 2.1049 and degrees of freedom of 15 and 782 for the numerator and denominator, respectively, we find evidence that past values of news sentiment contain information that is useful in predicting future values of MSFT stock returns.

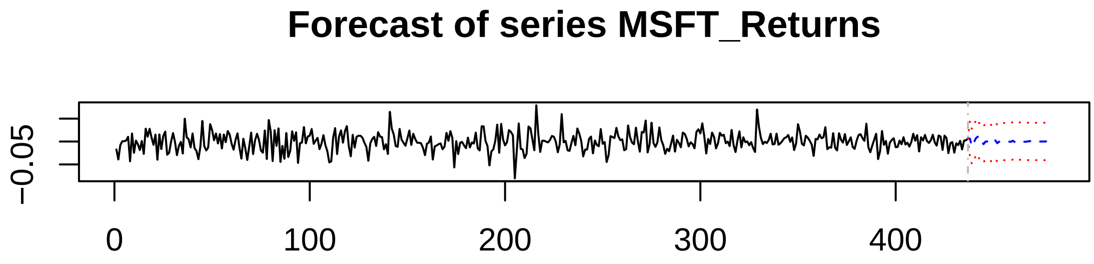
This finding calls for a reevaluation of the dynamic interplay between public sentiment, as captured by news, and subsequent investor behavior, as reflected in stock prices. It suggests that the informational content of news sentiment is indeed incorporated into market prices, albeit with a degree of complexity that may require further research to fully understand.

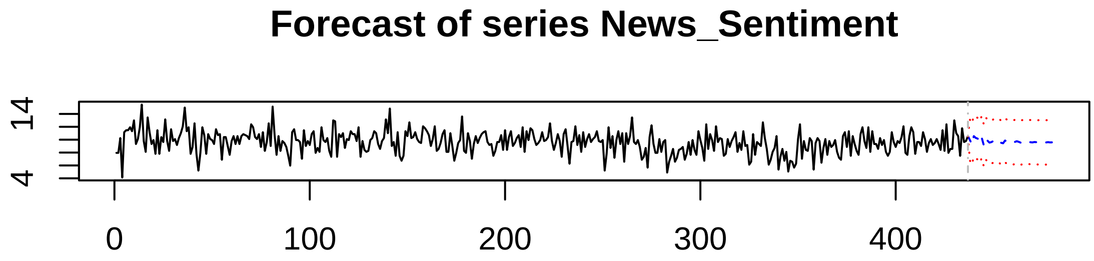
In light of this finding, our initial belief in the non-causal role of news sentiment may have been premature. This reinforces the need for a multifaceted approach to the analysis of financial time series, where causality may not always be apparent at first glance, but may emerge upon closer examination.

Forecasting

From a practical point of view, our VAR model may still be of some use when applied for forecasting purposes. To see if the inclusion of the news sentiment in our model has an improving effect on the models accuracy when predicting the price movements of MSFT we do some forecasting for our testing period where we include the news sentiment data of the testing period.

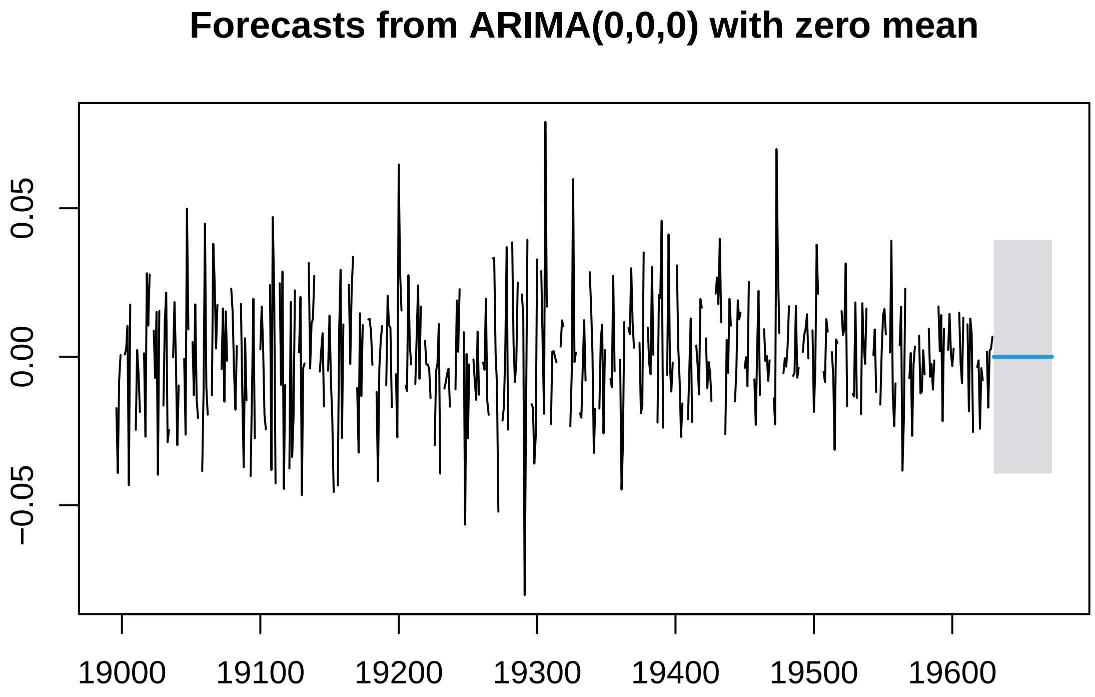
At the same time, we setup an ARMA model that is purely based on MSFTs previous price movements. It will serve as our benchmark.





```
## RMSE of VAR           0.03635417
```

```
## Series: MSFT_returns_train
## ARIMA(0,0,0) with zero mean
##
## sigma^2 = 0.0004022:  log likelihood = 1088.25
## AIC=-2174.51   AICc=-2174.5   BIC=-2170.06
```



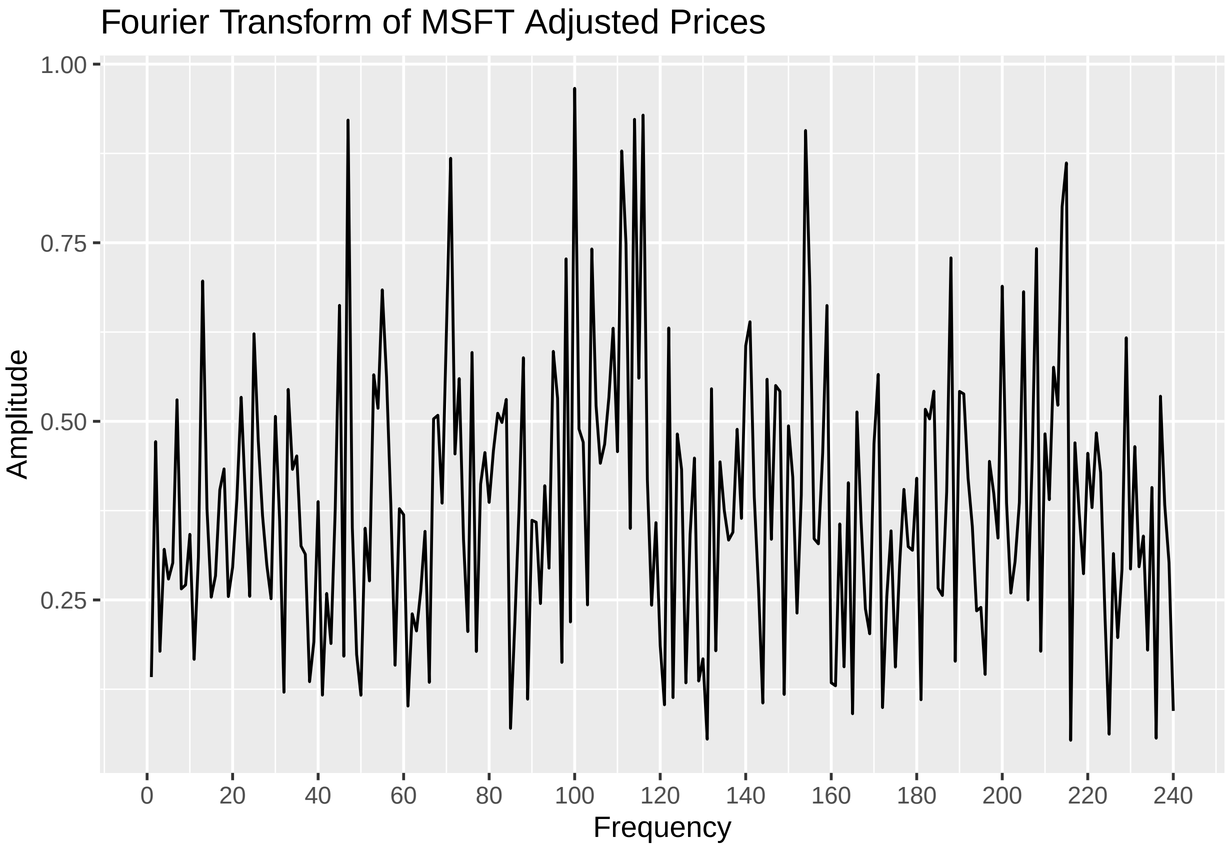
```
## RMSE of AR(2) Benchmark   0.01460572
```

We see that auto.arima has found that the order of the auto-regressive (AR) as well as the Moving Average (MA) parts are of order 0. Therefore, we essentially deal with a mean model as our benchmark. During forecasting, we also see that it uses a mean of 0 with a confidence band for all future values. This makes sense, as we deal with stationary data that commonly has a mean of 0.

While the VAR model reaches an RSME of 0.036 when compared to the truth values of our test set, the mean model benchmark reaches an RSME of 0.014. This means, our more sophisticated VAR model actually performs worse than the more simple benchmark that is purely based on the mean.

Seasonal Adjustment

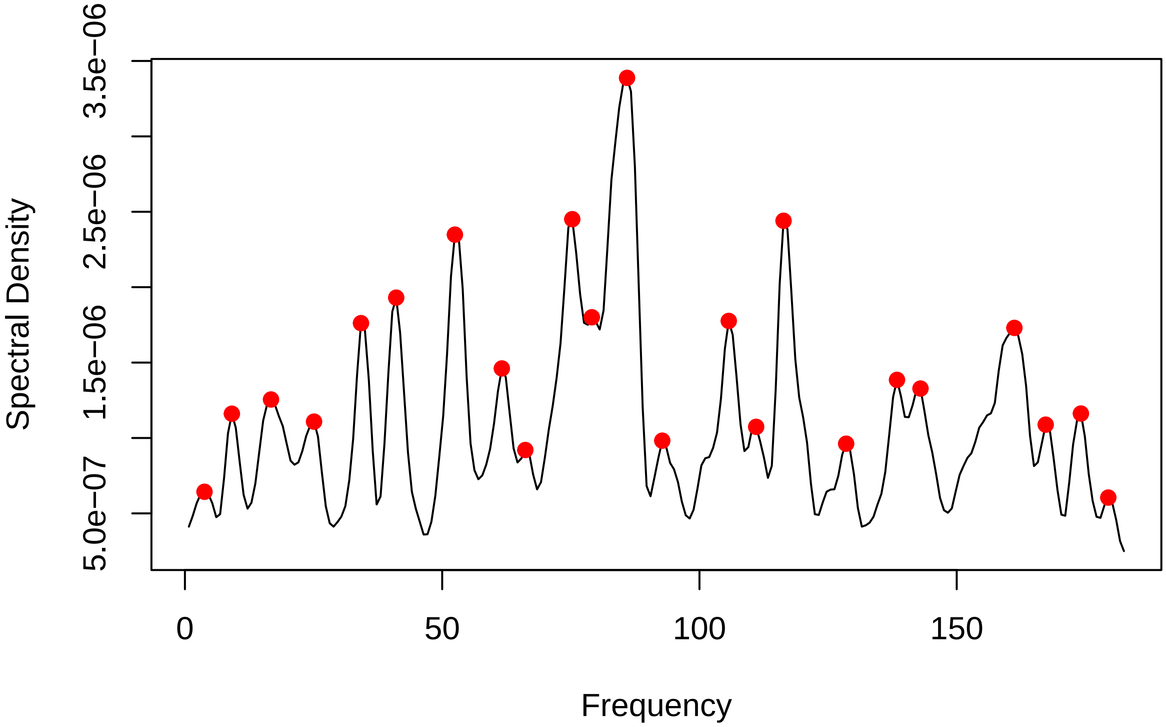
After many trials of extracting seasonality by using the STL and decompose method, we approached the problem using Fourier Transform due to the inconsistent nature of stock market data.



We notice in this plot the presence of multiple peaks suggesting that there are several cyclical patterns in the data. This can happen if there are multiple underlying seasonalities or other forms of cyclical behavior. But there is not a single, dominant low-frequency component that clearly stands out. This suggests that the data might not have a strong trend or annual seasonality, or it may be masked by the noise. The problem is if we attempt to model all these frequencies, there's a risk of overfitting your time series model to the noise rather than to actual, meaningful seasonal patterns.

Our Fourier analysis revealed multiple periodicities within the time series, suggesting complex seasonality. While including multiple Fourier terms can model these patterns, there is a risk of overfitting, especially if the data may contain noise masquerading as seasonal effects. To avoid this, we considered models that use a reduced set of Fourier terms, focusing only on the most significant frequencies. We also explored alternative methods, such as wavelet analysis, which can handle multiple levels of seasonality without assuming a fixed cycle length. These techniques allow us to capture the most important seasonal effects while reducing the risk of overfitting to the noise in the data.

Reevaluating seasonality and trend



## Model Selection

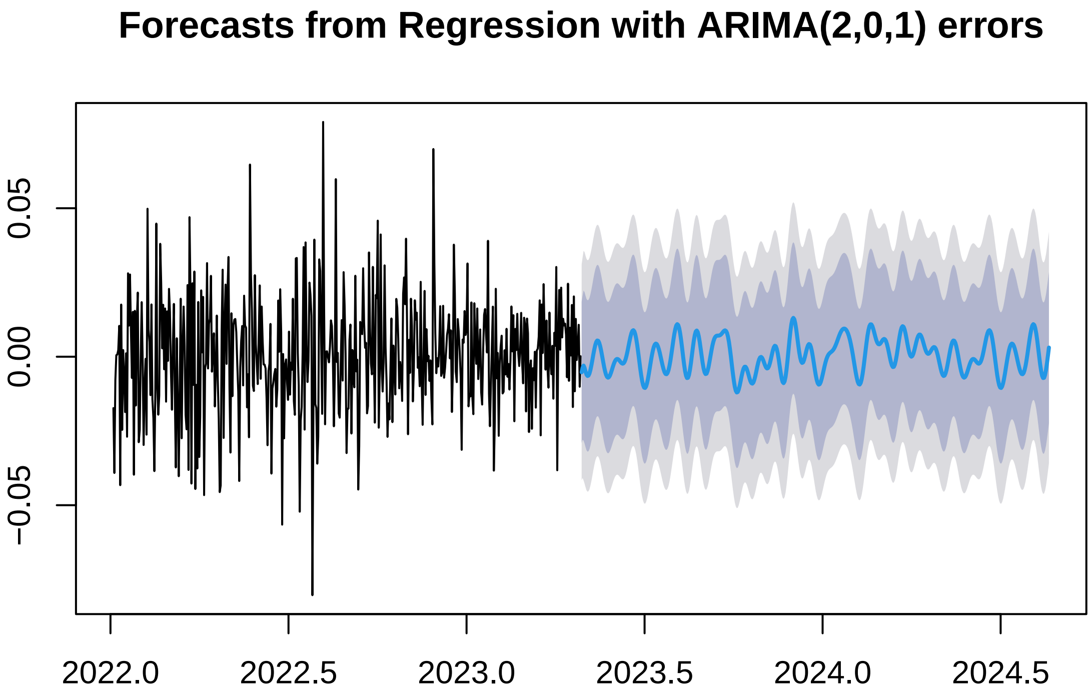
Based on the results of the Fourier analysis, we found it prudent to look for a model that would allow greater flexibility in capturing the identified seasonal patterns without overfitting. This led us to consider the ARIMA model, which can be augmented with Fourier terms to account for the complex seasonality observed. The ARIMA model was chosen for its ability to model non-seasonal and seasonal lags of the differenced series, which makes it suitable for our data, which exhibit non-stationarity. In addition, the inclusion of Fourier terms allows us to incorporate the significant seasonal frequencies identified earlier, providing a robust approach to capturing the underlying patterns in the data.

```
## Series: MSFT_returns_ts
## Regression with ARIMA(2,0,1) errors
##
## Coefficients:
##          ar1      ar2      ma1  intercept  S1-365  C1-365  S2-365  C2-365
##       0.5636  -0.1755  -0.7764      6e-04  0.0015   6e-04  0.0019  -3e-04
## s.e.  0.0587   0.0483   0.0414      3e-04  0.0005   4e-04  0.0004   4e-04
##        S3-365  C3-365   S4-365  C4-365  S5-365   C5-365  S6-365  C6-365  S7-365
##       -0.0012   5e-04  -0.0012  -4e-04   1e-03  -0.0013   9e-04  -8e-04   9e-04
## s.e.   0.0004   5e-04   0.0005   5e-04   5e-04   0.0005   5e-04   5e-04   5e-04
##       C7-365  S8-365   C8-365  S9-365   C9-365  S10-365  C10-365  S11-365
##        1e-03   0e+00  -0.0013  0.0027  -0.0015   0.0013  -0.0015   -1e-04
## s.e.   5e-04   5e-04   0.0005  0.0005   0.0005   0.0005   0.0005    5e-04
##       C11-365  S12-365  C12-365  S13-365  C13-365  S14-365  C14-365  S15-365
##         1e-04   -5e-04  -0.0017  -0.0013    5e-04    2e-04   0.0015  -0.0018
## s.e.    5e-04    6e-04   0.0006   0.0006    6e-04    6e-04   0.0006   0.0006
##       C15-365  S16-365  C16-365  S17-365  C17-365  S18-365  C18-365  S19-365
##         2e-04    9e-04  -0.0016    1e-04   0.0015    8e-04  -0.0018   0.0014
## s.e.    6e-04    6e-04   0.0006    7e-04   0.0007    7e-04   0.0007   0.0007
##       C19-365  S20-365  C20-365  S21-365  C21-365  S22-365  C22-365  S23-365
##        0.0012   0.0015  -0.0012    6e-04   -8e-04   -1e-04   0.0022   -8e-04
## s.e.   0.0007   0.0007   0.0007    8e-04    8e-04    8e-04   0.0008    8e-04
##       C23-365
##        0.0011
## s.e.   0.0008
##
## sigma^2 = 0.0003438:  log likelihood = 1259.22
## AIC=-2416.45   AICc=-2404.06   BIC=-2203.58
##
## Training set error measures:
##                         ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -0.0001083872 0.01754861 0.01380148 111.7995 194.3905 0.6607137
##                      ACF1
## Training set -0.006497629
```

The best-fit model for the series is an ARIMA(2,0,1). This model includes autoregressive (AR) terms, moving average (MA) terms, and various regression coefficients for exogenous variables. The model's log-likelihood is 1259.22, and it exhibits a reasonable AIC value, which suggest a good balance between model fit and complexity. The BIC value is also indicative of a well-fitting model. The training set error measures indicate that the model performs well on the training data, with a small mean error (ME) and low root mean squared error (RMSE) and mean absolute error (MAE). However, the MPE and MAPE are relatively high, suggesting some room for improvement in capturing certain patterns in the data. The MASE is 0.6607, and the autocorrelation of residuals (ACF1) is -0.0065.

Forecasting with the ARIMA Model

With our chosen model, we proceeded to forecasting. Here, we assess the model's predictive capability and compare its performance against a simpler benchmark model.



Calculate RMSE:

```
## RMSE for Model with Fourier Terms:  0.01597139
```

Forecasting Discussion:

In the forecasting section, you compare the performance of the ARIMA model with a benchmark model. It would be beneficial to discuss the implications of these results in more detail, explaining why the benchmark might perform better and what this suggests about the predictability of the stock price.

Conclusion

In summary, our comprehensive analysis using an ARIMA model with Fourier terms suggests that while there is a discernible pattern in the data, it is not strongly influenced by news sentiment as originally hypothesized. The non-significant results of the Granger causality test, coupled with the model's performance metrics, indicate that Microsoft's stock price movements are likely driven by factors not captured by news sentiment from the selected subreddits. The model's predictive performance indicates an area for potential improvement over a simpler benchmark model.

Future work

For future research, we recommend investigating additional variables that may better explain the variance in Microsoft stock prices, such as economic indicators, market indices, or sentiment from a broader range of news sources. It may also be beneficial to incorporate machine learning models capable of capturing non-linear relationships within the data. In addition, expanding the dataset to include high-frequency trading data could provide more granular insights into the impact of news events. Continued refinement of the model, including exploration of non-traditional forms of seasonality, will be critical to improving the predictive power of our analysis.

Bibliography:

Agus Shuarsono, Auliya Aziza and Wara Pramesti (2017); Comparison of vector autoregressive (VAR) and vector error correction models (VECM) for index of ASEAN stock price. DOI 10.1063/1.5016666

Aptech, Introduction to the Fundamentals of Vector Autoregressive Models (2021). https://www.aptech.com/blog/introduction-to-the-fundamentals-of-vector-autoregressive-models/

Microsoft, 2022 a look back at a year of accelerating progress in AI. https://www.microsoft.com/en-us/research/blog/2022-a-look-back-at-a-year-of-accelerating-progress-in-ai/

Roni Bhowmik, Gouranga Chandra Debnath, Nitai Chandra Debnath, Shouyang Wang (2022); Emerging stock market reactions to shocks during various crisis periods 2022). DOI 10.1371/journal.pone.0272450