Potential Drivers of Stock Market: Social Media and Financial News Sentiment

Quoc Minh DUONG
Group 20
Department of Data Science
City University of Hong Kong
mduong2-c@my.cityu.edu.hk

Hai Dang NGUYEN
Group 20
Department of Data Science
City University of Hong Kong
hdnguyen4-c@my.cityu.edu.hk

Thi Nguyet DUONG
Group 20
Department of Data Science
City University of Hong Kong
tnduong2-c@my.cityu.edu.hk

Abstract

This study investigates how sentiment from financial news and social media influences stock markets. Using tweets and financial news in 2020, daily sentiment values were calculated, and then their impacts on S&P 500 and NASDAQ-100 prices were tested via Granger causality, robust OLS regression with mediation analysis, and ARIMAX models. Results show tweet sentiment consistently helps predict price movements. News sentiment partially affects prices indirectly through tweets, explaining 26% of its total effect. Findings reveal prices and tweet sentiment influence each other, highlighting financial news sentiment as a stronger driver in total, but weaker in directly affecting stock market. This combined approach helps explain dynamics between sentiments and stock market beyond forecasting.

1 Introduction

In today's digital era, investors are constantly and widely exposed to social media platforms and financial news outlets, which serve as significant sources of information and can potentially influence their decisions and behaviors. Sentiment, the collective opinion expressed within these platforms, is thus considered a potentially valuable indicator of investor mood and market expectations. However, research often studies news or social media separately, emphasizing on using sentiment for prediction. Therefore, this study aims to provide a more integrated understanding of how these two sentiment sources together influence overall market performance, rather than solely focusing on forecasting. To achieve this aim, this research seeks to answer the following key questions:

1. How do social media sentiment and financial news sentiment relate to changes in stock prices, and what is the nature (including predictive power, directness and mediated pathways) of these relationships?
2. How does the strength of the relationship between financial news sentiment and stock prices compare to that of social media sentiment and stock prices?

By answering these questions, this study provides an explanatory analysis of how social media and financial news sentiment could potentially be drivers of stock prices.

2 Background

Numerous studies have investigated how sentiment derived from social media and financial news can be applied to forecast stock market movements.

A number of studies have investigated stock market behavior in relation to financial news. One of these studies sought to measure the impact of news sentiment on stock prices through sentiment analysis and models such as logistic regression, as well as neural networks, followed by Recurrent Neural Networks (RNN) for forecasting actual prices [4]. Another developed a model that integrates structured financial information with unstructured news content for forecasting stock price volatility, with some notable correlations found between volatility score and sentiment for different periods [6]. Most notably, large-scale financial news collection is presented in [3], together with LSTM-based deep learning utilising Focal Calibration Loss, yielding greater reliability in predictions through increased accuracy and calibration.

Social media platforms have emerged as rich sources of knowledge about investor sentiment and its correlation with stock price movements. A novel 10,000 annotated StockTwits tweet dataset with binary sentiments (bullish/bearish) as well as 12 fine-grained emotion types is proposed in [2]. To investigate the impact of these emotions on financial markets, the authors used a Temporal Attention LSTM model that combined sentiment, emotion, and stock price index features within an integrated multivariate time series forecasting structure. Another study used Yahoo! Finance message boards as input for aspect-based sentiment analysis coupled with Linear SVM classifiers for predicting stock direction of movement. In the study, sentiment features were found to yield results superior to price-only baselines, notably for low-predictability stocks [5]. Big data techniques with PySpark and MLlib are utilized on 10-year-long historical stock data alongside social media data in [1]. The study investigated a number of machine learning models, including Linear Regression, Random Forest, Generalized Linear Regression (GLR) and Decision Tree to forecast changes in stock prices. The most accurate of these was Generalized Linear Regression, which in certain situations reached up to 97% accuracy.

Although previous studies established various models that involve sentiment from either financial news or social media, such work is mostly concerned with predicting market outcomes instead of uncovering the mechanisms for their impact. Moreover, prior research tends to analyze each source of sentiment in isolation, with the interactive possibilities between them not being considered. The current study adopts a different approach by shifting the focus from prediction to explanation. With both financial news and social media sentiment treated in an integrated analytical framework, this study not only explores the impact of sentiment on the stock index, but also, potentially, whether a source of sentiment can affect market outcomes through the other. The integrated and explanation-aware approach of this work provides a fuller picture of how digital sentiment courses through multiple pathways and jointly influences the behavior of investors in the financial market.

3 Method

3.1 Dataset

This study leverages two externally available datasets - StockEmotions and FinSen—which capture sentiment signals from social media and financial news, respectively. Both of them are temporally aligned for the observation period of 2020, corresponding with the period for which stock index performance is studied.

The StockEmotions dataset comprises 10,000 English tweets from the StockTwits platform in 2020. Each tweet is manually labeled as either bullish (assigned a value of 1) or bearish (assigned –1). To generate a daily sentiment score, the total sentiment score is aggregated for each day and divided by the total number of tweets, producing a Tweet Sentiment Score as follows:

$$\text{Tweet Sentiment Score} = \frac{\text{Total Daily Sentiment Score}}{\text{Tweet Count}}$$

This score reflects the average polarity of investor sentiment expressed on social media each day.

The FinSen dataset contains over 160,000 financial news from 197 countries from 2007 through 2023. For the purpose of this study, articles from 2020 and only from the US are used in order to ensure alignment with the StockEmotions dataset. Sentiment analysis is performed using FinBERT, which is a transformer-based fine-tuned for financial text classification. The term SAggd is used by convention as defined in [3], where it refers to the aggregated daily sentiment score derived from financial news using FinBERT [7]. It is calculated using the below equation, where N_d refers to the total number of articles on that given day.

$$SAgg_d = \frac{1}{N_d} \sum_{i=1}^{N_d} ((-1 \cdot P_{\text{neg}}) + (0 \cdot P_{\text{neu}}) + (1 \cdot P_{\text{pos}}))$$

Since the independent and mediating variables are aggregates, we also use historical data of two stock indices: S&P 500 and NASDAQ-100, which are aggregates of many individual stocks in the US. After integrating all the data sources, the final dataset has 252 data points, each representing the daily aggregated variables in 2020.

3.2 Metrics

This study examines the influence of sentiment on market behavior by defining a structured set of dependent, independent, and mediating variables. The dependent variable is the daily closing price of two key stock indices: the S&P 500 and the NASDAQ-100. These measures capture broad market dynamics and serve as the primary outcome variables in assessing the impact of sentiment on financial performance.

The independent variables consist of two sentiment measures: the daily financial news sentiment score, referred to as SAggd, and the Tweet Sentiment Score derived from social media posts. SAggd reflects aggregated sentiment derived from financial news processed through FinBERT, while the Tweet Sentiment Score captures the average daily sentiment expressed on StockTwits, adjusted for the total number of tweets posted.

To investigate whether sentiment exerts influence through indirect pathways or supported by mediators, the study includes several mediating variables. The proposed causal path is indicated in Figure 1. Mediating variables consist of the number of financial news articles published each day (Article Count), as well as the number of articles identified as either positive (Bullish Count) or negative (Bearish Count) in tone. In addition, the analysis considers whether sentiment from financial news—captured by SAggd—might indirectly shape social media sentiment, based on the reasonable assumption that individuals often respond to news content in their online posts. SAggd is therefore also considered as a potential mediator of sentiment on tweets. This setup makes it possible for both the direct and indirect channels through which sentiment is influencing market indices to be examined.

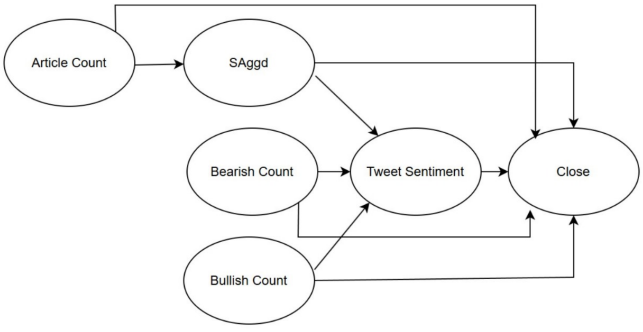

Figure 1: Proposed Causality Path

3.3 Model

To better understand how sentiment affects the stock market over time, this study uses three complementary methods: Granger causality tests, OLS regression with HAC standard errors, and ARIMAX modeling.

Granger Causality test: Granger causality is used to test whether changes in sentiment can help predict future movements in the market. Similar to the FinSen Study, this paper applies Granger tests to see if financial news sentiment (SAggd) can support predicting stock behavior. However, unlike previous work, this study focuses on the stock prices instead of stock volatility and also includes social media sentiment, offering a broader perspective on how different sources of sentiment might influence market behavior.

OLS regression with HAC standard errors: To gain an increased understanding about the size and direction of these effects, the study also employs OLS regression but with an important adjustment. Because market data are time-based and often autocorrelated, standard OLS can lead to biased results. To address this, the model uses Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors with the Newey–West estimator, which helps correct for issues like autocorrelation and changing variance over time. The results then become more reliable and the p-values become more valid.

ARIMAX: Finally, we employs ARIMAX, a time-series framework that can capture patterns in historical stock prices and integrate exogenous inputs specified above. This model aligns well with our time-series data and research purpose. By using ARIMAX, we can observe whether sentiment can provide additional insight beyond what is revealed from the market's past behavior.

4 Results

4.1 Granger Causality test

We first evaluate the stationarity of daily tweet and financial news sentiment scores and also all price variables from S&P 500 and NASDAQ-100 (High, Low, Open, Close). This is performed by using the Augmented Dickey-Fuller (ADF) test, and differencing is applied on non-stationary variables. Following this preprocessing, we conducted Granger Causality tests to examine the predictive relationships between sentiment variables and stock market prices. These tests were performed using a lag length of 7 and a significance level of $\alpha = 0.05$. The analysis specifically assessed potential statistical causality in both directions: testing whether past price information Granger-causes sentiment, and conversely, whether past sentiment information Granger-causes prices.

A surprising finding is that the test results for both S&P 500 and NASDAQ-100 are the same in both directions. From Table 1, it is suggested that lagged values of financial news sentiment (SAggd) do not provide statistically significant additional predictive power for stock prices beyond the information already contained in the lagged values of stock prices. In contrast, tweet sentiment shows significant forecasting capability, where it can help improve prediction at all 7 lags for all price variables, except Close. Table 2 indicates that High, Low, and Close have a strong predictive relationship with Tweet Sentiment, suggesting market performance influences social media discussions. This suggests a bidirectional relationship between stock prices and social media sentiment, where one helps predict the other and vice versa. Meanwhile, stock prices do not show any significant predictive power for financial news sentiment.

X \ Y	High	Low	Close	Open
SAgged Tweet Sentiment	No significant lags All 7 lags	No significant lags All 7 lags	No significant lags Significant at lag 1	No significant lags All 7 lags

Table 1: Granger Causality Tests for Sentiments -> Stock Prices

X \ Y	SAgged	Tweet Sentiment
High	No significant lags	All 7 lags
Low	No significant lags	All 7 lags
Close	No significant lags	All 7 lags
Open	No significant lags	No significant lags

Table 2: Granger Causality Tests for Stock Prices -> Sentiments

4.2 OLS Regression with HAC Standard Errors

Before fitting into the model, independent variables and mediating variables are first preprocessed by applying a moving window. More specifically, SAggd and tweet sentiment are smoothed using a rolling window of 30 days, while all other mediating variables are applied a 60-day window. This smoothing technique is used primarily to reduce the impact of short-term volatility and high-frequency noise inherent in daily data, while aligning with the fact that the influence of sentiment comes from both the present and past values. A problem with directly applying OLS regression is that time series data can exhibit heteroskedasticity (non-constant error variance) and autocorrelation (correlation between errors over time) in the residuals of an OLS regression. These violations of the standard OLS assumptions would render the traditional standard errors biased and inconsistent, leading to unreliable hypothesis testing and inference about the significance of the predictors. Therefore, after fitting the OLS model, Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors, specifically the Newey-West estimator, are employed to provide robust and reliable estimates of the standard errors that account for the potential presence of both heteroskedasticity and autocorrelation in the model residuals.

The R-squared values for NASDAQ-100 and S&P 500 are 0.890 and 0.779, respectively, showing that the exogenous variables explain the variance in the closing prices of the indices relatively well. As expected, from Table 3 and Table 4, it is shown that SAggd and Tweet Sentiment are no longer significant after accounting for autocorrelation and heteroskedasticity. Regardless, meaningful insights can be provided from the coefficients of the variables. For both the NASDAQ-100 and S&P 500, the signs of the coefficients are consistent with expectations: Article Count, Bullish Count, SAggd, and Tweet Sentiment are positively associated with higher stock prices, while Bearish Count shows a negative association. For a unit increase in the smoothed financial news sentiment, the closing price of NASDAQ-100 and S&P 500 rises by $1088.77 and $313.838, respectively. Similarly, a unit increase in the smoothed tweet sentiment results in a growth of $1053.1206 for NASDAQ-100 and $268.7875 for S&P 500. Comparing the standardized coefficients, SAggd appears to have a slightly larger estimated total effect than Tweet Sentiment on both indices, but a smaller direct effect.

Variables	Unstandardized	Standardized	p-value	p-value with HAC
SAggd	1088.77	0.1132	0.001	0.075
Tweet Sentiment	1053.1206	0.1455	0.001	0.122
Article Count	767.647	0.8747	0	0
Bullish Count	145.7725	0.5234	0	0
Bearish Count	-168.1644	-0.5924	0	0

Table 3: OLS Regression Results for NASDAQ-100

Variables	Unstandardized	Standardized	p-value	p-value with HAC
SAggd	313.838	0.1539	0.001	0.092
Tweet Sentiment	268.7857	0.1751	0.004	0.195
Article Count	116.2204	0.6244	0	0
Bullish Count	30.7846	0.5211	0	0
Bearish Count	-42.6839	-0.7089	0	0

Table 4: OLS Regression Results for S&P 500

Based on the decomposed effects presented in Table 5 and Table 6, the analysis of mediation pathways reveals that for most variables, the indirect effects are relatively small in magnitude compared to their direct and total effects. However, SAggd appears to be an exception, exhibiting a notable indirect effect in both indices. This indirect effect, under the assumption that financial news sentiments affect tweet sentiments (people tweets after reading the news), accounts for approximately 27.94% of SAggd's total effect on NASDAQ-100 stock prices and approximately 25.55% of its total effect on S&P 500 stock prices.

Variable	Direct Effect	Indirect Effect	Total Effect
SAggd	0.1132	0.0439	0.1571
Tweet Sentiment	0.1455	0	0.1455
Article Count	0.8747	-0.0213	0.8535
Bullish Count	0.5234	0.0601	0.5834
Bearish Count	-0.5924	-0.1058	-0.6981

Table 5: Decomposed Effects of Variables on NASDAQ-100 Stock Prices

Variable	Direct Effect	Indirect Effect	Total Effect
SAggd	0.1539	0.0528	0.2067
Tweet Sentiment	0.1751	0	0.1751
Article Count	0.6244	-0.0280	0.5964
Bullish Count	0.5211	0.0723	0.5934
Bearish Count	-0.7089	-0.1273	-0.8362

Table 6: Decomposed Effects of Variables on S&P 500 Stock Prices

4.3 ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables)

Before fitting the ARIMAX model, the dependent variable is differenced once ($d = 1$) to make it stationary. Based on that, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are utilized to analyze the time series data and determine the appropriate autoregressive (AR) and moving average (MA) orders based on its autocorrelation structure. The result indicates $p = 1$ and $q = 1$ are the best orders for the model. The equations acquired are shown in Equation 1 and Equation 2.

$$\begin{aligned}\Delta\text{Close}_{\text{NASDAQ},t} &= -0.6045\Delta\text{Close}_{\text{NASDAQ},t-1} + 0.2706\epsilon_{\text{NASDAQ},t-1} \\ &+ 58.6291\text{SAggd}_t + 70.2598\text{tweet\_sentiment}_t - 1.0039\text{Article\_Count}_t \\ &+ 1.5285\text{bullish\_count}_t + 0.0184\text{bearish\_count}_t + \epsilon_{\text{NASDAQ},t}\end{aligned} \tag{1}$$

$$\begin{aligned}\Delta\text{Close}_{\text{S\&P500},t} &= -0.6142\Delta\text{Close}_{\text{S\&P500},t-1} + 0.2723\epsilon_{\text{S\&P500},t-1} \\ &+ 29.8934\text{SAggd}_t + 15.3797\text{tweet\_sentiment}_t - 0.8043\text{Article\_Count}_t \\ &+ 0.2977\text{bullish\_count}_t + 0.0048\text{bearish\_count}_t + \epsilon_{\text{S\&P500},t}\end{aligned} \tag{2}$$

From the equations, most of the coefficients align with the results from OLS regression, where SAggd, Tweet Sentiment, Bullish Count have positive coefficients, suggesting that increases in these factors are associated with positive movements in closing prices. Conversely, different from OLS regression, Bearish Count has positive coefficients and Article Count shows a negative coefficient in both models, indicating that a rise in the number of articles is associated with a decrease in closing price, but a rise in the number of negative articles facilitates the increase in closing price.

Based on Tables 7 and 8, an increase of one unit in the daily sentiment score of financial news is associated with an estimated increase of $58.6291 in the daily change of the NASDAQ closing price and $29.8934 in the daily change of the S&P 500 closing price. The corresponding numbers regarding the daily tweet sentiment scores are $70.2598 for NASDAQ-100 and $15.3797 for S&P 500. Both SAggd and Tweet Sentiment demonstrate significant or marginally significant relationships with daily changes in closing prices across the two indices. SAggd shows a highly statistically significant positive coefficient for the NASDAQ-100 (p=0.001) and a significant positive coefficient for the S&P 500 (p=0.048), indicating its consistent importance. Tweet Sentiment is also a significant positive predictor for the S&P 500 (p=0.007) and is marginally significant for the NASDAQ-100 (p=0.054). For S&P 500, Tweet Sentiment imposes a stronger direct effect than SAggd, aligning with results from OLS regression, but it is the opposite for NASDAQ-100 where SAggd has a slightly stronger impact.

Variable	Unstandardized Coeff.	Standardized Coeff.	p-value
SAggd	29.8934	0.0403	0.001
Tweet Sentiment	15.3797	0.0305	0.054
Article Count	-0.8043	-0.0139	0.404
Bullish Count	0.2977	0.0173	0.170
Bearish Count	0.0048	0.0003	0.979

Table 7: ARIMAX Exogenous Variable Coefficients for NASDAQ-100

Variable	Unstandardized Coeff.	Standardized Coeff.	p-value
SAggd	58.6291	0.0176	0.048
Tweet Sentiment	70.2598	0.0303	0.007
Article Count	-1.0039	-0.0039	0.710
Bullish Count	1.5285	0.0185	0.027
Bearish Count	0.0184	0.0007	0.975

Table 8: ARIMAX Exogenous Variable Coefficients for S&P 500

5 Conclusion

By leveraging multiple methods, evaluating and comparing their results, this study was able to dive deep into the impact significance, strength, direction, and mechanism of social media and financial news sentiment on the overall stock market prices. First, it is found that tweet sentiment potentially has a bidirectional predictive relationship with stock behavior, where one helps predict the other. Financial news sentiment, on the other hand, presents no significant lags, suggesting little forecasting capability for stock prices. Secondly, the influence of sentiments on stock prices is mostly intuitive. Generally, higher positivity in tweets and financial articles is associated with overall higher stock prices; however, inconsistencies are found, suggesting a more complex interaction to be explored. In addition, around 26% the impact of financial news sentiment on stock prices was found to be indirect through tweet sentiment. Although the OLS regression results regard sentiments as insignificant after accounting for autocorrelation and heteroskedasticity, from the ARIMAX model, sentiments from tweets and financial news indicate significant or marginally significant relationships with overall daily changes in the stock market. Finally, the results generally suggest that while financial news sentiment may have a larger total effect on stock prices, its direct influence appears smaller compared to sentiment derived from tweets.

While this study provides valuable insights into the relationship between sentiments and stock market behavior, it is subject to certain limitations. One limitation of this study is the usage of a dataset and stock indices exclusively from the US. Consequently, the generalizability of our findings to other international markets may be limited. Another limitation stems from the unit of analysis, which are aggregates of individual stocks (NASDAQ and S&P 500 indices). The dynamics observed at the index level, being a composite of many stocks, may not fully capture or generalize to the specific price movements and sentiment relationships of individual equities. Furthermore, while a longitudinal framework was utilized to examine temporal relationships, applying time-series data to OLS regression, even with HAC might cause potential problems such as spurious relationship, requiring careful interpretation of the results. In addition, the analysis was restricted to only two subjects. This matter, especially for explanatory analysis, constraints the ability to draw conclusions generalizable to other indices or equities. This limitation is due to the lack of large-scale labelled tweet and financial news data for individual stocks. The financial news data was not labelled which stock it refers to, while the tweet data labels are mostly some well-known stocks only. With access to a more diverse and comprehensively labeled dataset encompassing both tweet and financial article sentiment for numerous individual stocks, a promising future research direction for explanatory analysis would be to leverage panel data methodologies and models such as Fixed-Effect, Random Effects, or even dynamic panel models like Generalized Method of Moments (GMM). While acknowledging these constraints, this research provides initial meaningful findings that lay the groundwork for future studies to investigate the explanatory mechanisms linking sentiment to stock market behaviors.

References

[1] Mazhar Javed Awan, Mohd Shafry Mohd Rahim, Haitham Nobanee, Ashna Munawar, Awais Yasin, and Azlan Mohd Zain. Social media and stock market prediction: A big data approach. *Expert Systems with Applications*, 2021.

[2] Jean Lee, Hoyoul Luis Youn, Josiah Poon, and Soyeon Caren Han. Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series. *arXiv:2301.09279 [cs]*, 2023.

[3] Wenhao Liang, Zhengyang Li, and Weitong Chen. Enhancing financial market predictions: Causality-driven feature selection. *IEEE Xplore*, 2024.

[4] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C. Anastasiu. Stock price prediction using news sentiment analysis. *IEEE Xplore*, 2019.

[5] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 2015.

[6] Jia-Lang Seng and Hsiao-Fang Yang. The association between stock price volatility and financial news – a sentiment analysis approach. *IEEE Xplore*, 2017.

[7] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv:2006.08097 [cs]*, 2020.

Appendix

1. Granger Causality Test

The Granger Causality test is used to determine whether one time series provides statistically significant information in forecasting another. In this study, it helps assess whether sentiment scores (such as *SAggd* or *Tweet Sentiment Score*) can predict movements in stock indices.

Let X_t and Y_t be two stationary time series. The null hypothesis states that X_t does not Granger-cause Y_t. This is tested using the following regression model:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j X_{t-j} + \varepsilon_t$$

If the coefficients β_j are jointly statistically significant, we reject the null hypothesis and conclude that X Granger-causes Y.

2. Newey–West HAC Regression

Ordinary Least Squares (OLS) regression assumes that the error terms are independently and identically distributed. However, time series data often suffer from autocorrelation and heteroskedasticity, which violate these assumptions. The Newey–West estimator adjusts the standard errors of OLS coefficients to remain consistent under these conditions.

For a linear regression model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

The Newey–West estimator for the variance-covariance matrix of the coefficients is given by:

$$\widehat{V}_{\text{HAC}} = \hat{\sigma}^2 \left(\sum_{t=1}^{T} x_t x_t' + \sum_{\ell=1}^{L} w_\ell \sum_{t=\ell+1}^{T} \left(x_t x_{t-\ell}' + x_{t-\ell} x_t' \right) \right)$$

where $w_\ell = 1 - \frac{\ell}{L+1}$ is the Bartlett kernel and L is the selected lag length. Following common practice, the lag length L for the Newey–West estimator was selected using the data-dependent heuristic $L = \lfloor 4(T/100)^{2/9} \rfloor$, where T is the sample size.

3. ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables)

The ARIMAX model is a time series model that incorporates both autoregressive and moving average components, along with external (exogenous) variables. This allows the model to explain the behavior of a time-dependent variable based on both its own history and the influence of outside predictors such as sentiment scores.

The general form of the ARIMAX model is:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=0}^{r} \gamma_k x_{t-k} + \varepsilon_t$$

where:

- y_t is the dependent variable (e.g., stock index value),
- x_t is an exogenous input (e.g., sentiment score),
- ε_t is the white noise error term,
- p is the autoregressive (AR) order,
- d is the differencing order (if applicable),

- q is the moving average (MA) order.

This model is particularly useful for identifying whether sentiment provides additional predictive power beyond the trends already present in the stock data itself. The coefficients are valuable for determining the strength and direction of the relationship between sentiments and stock data. Visualization for ACF and PACF, as well as details for the ARIMAX models are presented in the figures below.

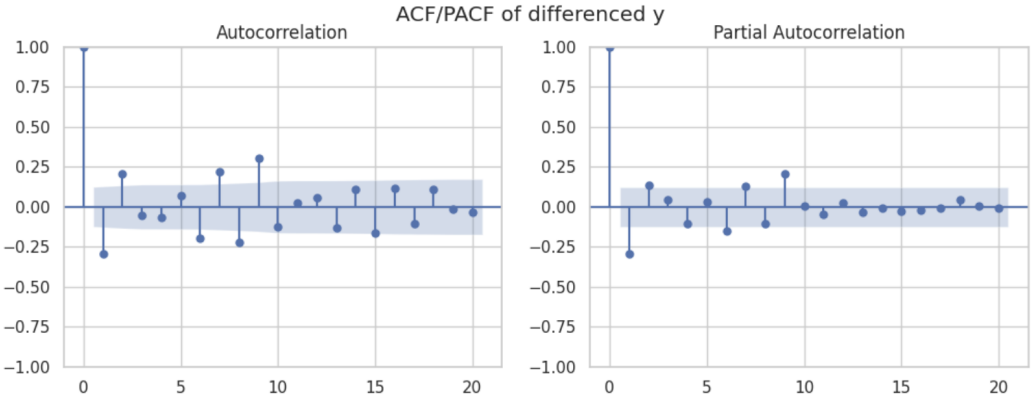

Figure 2: ACF and PACF of NASDAQ-100 ($d = 1$)

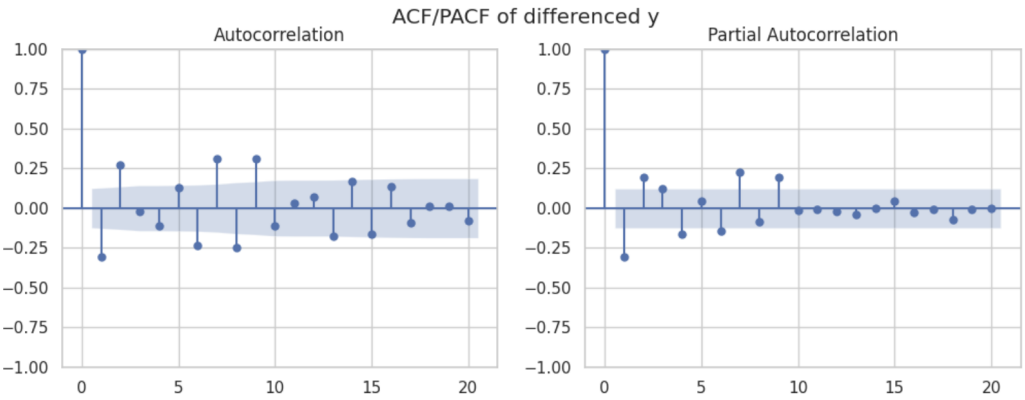

Figure 3: ACF and PACF of S&P 500 ($d = 1$)

Dep. Variable:	Close	No. Observations:	252
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1661.068
Date:	Sun, 04 May 2025	AIC	3338.136
Time:	15:44:10	BIC	3366.340
Sample:	0	HQIC	3349.486
	- 252		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
SAggd	58.6291	29.654	1.977	0.048	0.508	116.750
tweet_sentiment	70.2598	26.056	2.696	0.007	19.190	121.330
Article_Count	-1.0039	2.695	-0.373	0.710	-6.286	4.278
bullish_count	1.5285	0.693	2.204	0.027	0.170	2.887
bearish_count	0.0184	0.590	0.031	0.975	-1.138	1.174
ar.L1	-0.6045	0.104	-5.811	0.000	-0.808	-0.401
ma.L1	0.2706	0.115	2.348	0.019	0.045	0.496
sigma2	3.263e+04	2333.073	13.984	0.000	2.81e+04	3.72e+04

Ljung-Box (L1) (Q):	0.13	Jarque-Bera (JB):	72.51
Prob(Q):	0.72	Prob(JB):	0.00
Heteroskedasticity (H):	0.75	Skew:	-0.68
Prob(H) (two-sided):	0.20	Kurtosis:	5.26

Figure 4: Detailed results of ARIMAX for NASDAQ-100

Dep. Variable:	Close	No. Observations:	252
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1354.416
Date:	Sun, 04 May 2025	AIC	2724.831
Time:	15:44:32	BIC	2753.035
Sample:	0	HQIC	2736.181
	- 252		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
SAggd	29.8934	8.969	3.333	0.001	12.314	47.473
tweet_sentiment	15.3797	7.976	1.928	0.054	-0.253	31.012
Article_Count	-0.8043	0.964	-0.834	0.404	-2.694	1.085
bullish_count	0.2977	0.217	1.372	0.170	-0.128	0.723
bearish_count	0.0048	0.180	0.026	0.979	-0.349	0.358
ar.L1	-0.6142	0.086	-7.103	0.000	-0.784	-0.445
ma.L1	0.2723	0.096	2.837	0.005	0.084	0.460
sigma2	2844.3835	166.193	17.115	0.000	2518.652	3170.115

Ljung-Box (L1) (Q):	0.70	Jarque-Bera (JB):	299.45
Prob(Q):	0.40	Prob(JB):	0.00
Heteroskedasticity (H):	0.31	Skew:	-1.07
Prob(H) (two-sided):	0.00	Kurtosis:	7.90

Figure 5: Detailed results of ARIMAX for S&P 500

4. Source Code

Our data and source code for visualization, analysis, and modeling can be found in this GitHub Repository.