University of Science and Technology of Hanoi
Department of ICT

News Sentiment Analysis and Stock Price Impact

Vũ Hải Thiên Long
22BI13268
LongVHT.22BI13268@usth.edu.vn

Lưu Linh Ly
22BI13269
LyLL.22BI13269@usth.edu.vn

Lê Viết Hoàng Lâm
22BI13235
LamLVH.22BI13235@usth.edu.vn

ABSTRACT

This report explores the integration of sentiment analysis and machine learning techniques, specifically Long Short-Term Memory (LSTM) networks, to enhance stock price prediction. Traditional methods often overlook the significant impact of market sentiment driven by news events, which can lead to sudden price fluctuations. By analyzing historical stock prices alongside sentiment scores derived from news articles, this study aims to improve forecasting accuracy. We focus on key features such as adjusted close prices, changes in stock prices, sentiment scores, news counts, and 7-day moving averages. The results demonstrate that incorporating sentiment data into LSTM models significantly enhances predictive performance, offering valuable insights for investors and financial analysts in making informed decisions.

TABLE OF CONTENT

1. Introduction

Predicting stock prices poses a significant challenge in finance, as accurate forecasts are essential for investors and analysts. While traditional models focus on historical trends and economic indicators, they often overlook the crucial impact of market sentiment. Events in the news and public opinion can significantly influence stock prices, making real-time sentiment analysis a valuable tool for improving predictions.

Traditional methods, such as time series and technical analysis, rely on past data but fail to account for emotional responses to news, leading to volatility. To address this gap, data science introduces sentiment analysis and machine learning, particularly recurrent neural networks like LSTMs. These approaches leverage natural language processing to quantify sentiment from news and social media, integrating it into stock prediction models. This report explores how incorporating sentiment data into LSTM models—through features like adjusted close prices, stock changes, sentiment scores, news counts, and moving averages—can enhance forecasting accuracy and support better financial decision-making in volatile markets.

2. Background

Predicting stock prices is a central challenge in finance, where accurate forecasts are crucial for investors, analysts, and financial institutions to make informed decisions. Stock prices are influenced by numerous factors, including historical trends and economic indicators, but traditional models relying solely on these data sources often overlook the impact of market sentiment. News events and public opinion can dramatically affect stock prices, and capturing these shifts in real-time provides a significant advantage.

Traditional methods, such as time series and technical or fundamental analysis, rely on past prices, volume, and financial statements to predict trends. While effective, these approaches lack the ability to account for emotional and psychological reactions to news events, which can lead to sudden and unexpected volatility in stock prices. The need for methods that integrate both quantitative data and sentiment information has become increasingly clear.

Data science addresses this by introducing sentiment analysis and machine learning. Sentiment analysis uses natural language processing (NLP) to quantify public opinion from news and social media, generating sentiment scores that can enhance stock prediction models. Machine learning, especially with recurrent neural networks like LSTMs, captures complex relationships between price data and sentiment over time, allowing for more accurate and responsive forecasts. This work explores integrating sentiment data into LSTM-based models, aiming to enhance stock price prediction through sentiment-aware analysis for improved financial decision-making.

3. Methodology

3.1 System design

Following system design is proposed in this project to classify news articles sentiment for understanding stock price impact.

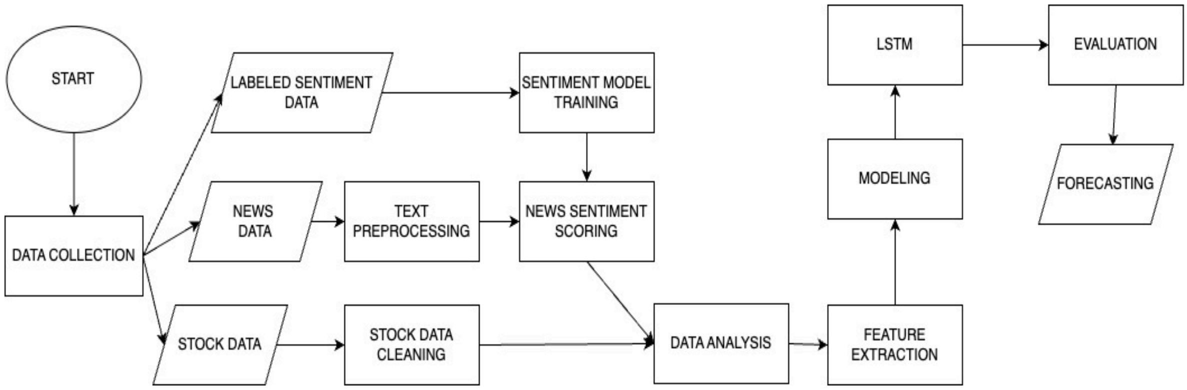


Figure 1. System Design

This design can be divided into two phases: The first phase focuses on data preprocessing, assigning sentiment scores, and selecting the most relevant features for the next phase. The second phase is the modeling phase, where the processed data is used to train models that predict stock prices with unseen data.

3.2 Data Preparation

3.2.1 Data Collection

We analyzed news and stock prices for Microsoft Corp. (MSFT), NVIDIA Corp. (NVDA), and Intel Corp. (INTC) to study how news sentiment affects stock price movements across different volatility levels, as indicated by their beta values [1]:

- Microsoft Corp. (MSFT): **0.96**
- Intel Corp. (INTC): **1.5**
- NVIDIA Corp. (NVDA): **2.42**

Beta measures a stock's volatility relative to the market (S&P 500 beta = 1). A beta above 1 signals higher volatility, as with NVIDIA, which shows greater sensitivity to market shifts, increasing both risk and potential returns. In contrast, Microsoft's beta of 0.96 indicates more stable movements aligned with the market.

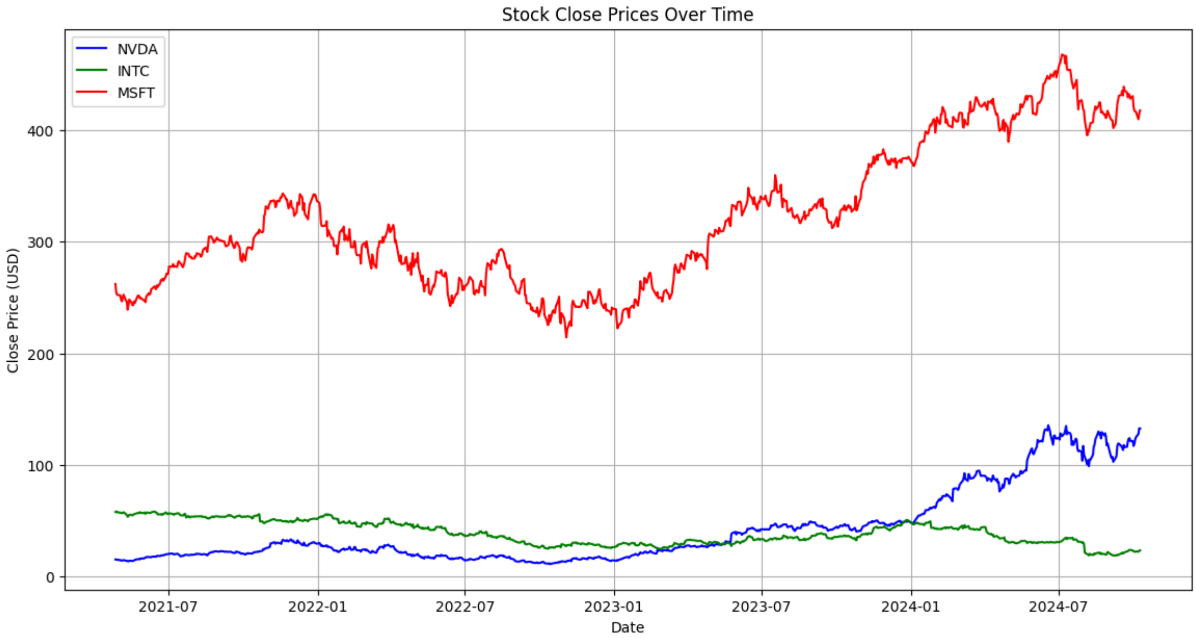


Figure 2. Stock Close Price across 3 stocks

The news and stock prices used to train were collected from April 27, 2021, to October 9, 2024 and the data we collected to test the model were from October 10th to October 24, 2024. News data was sourced from the Polygon.io API, while stock prices were obtained from Yahoo Finance. For stocks, we gathered data on the *Adjusted Close*, *Date*, Open*, High, Low* and *Volume*. For news, we collected the *Title*, *Content*, *Publisher*, and *Published Date*.

	title	content	publisher	published_date
0	Intel's Granite Rapids Could Drive a Major Reb...	Intel's data center business has struggled in ...	The Motley Fool	2024-10-05T09:50:00Z
1	Marjorie Taylor Greene Continues Adding to Her...	Marjorie Taylor Greene, a Republican represent...	The Motley Fool	2024-10-03T11:15:00Z
2	The Shocking Truth About Intel's 18A Node and ...	Intel's upcoming 18A manufacturing node is cru...	The Motley Fool	2024-10-01T11:45:00Z
3	What's the Outlook on Spices?	Intel is drawing interest from Qualcomm and Ap...	The Motley Fool	2024-09-30T20:56:00Z
4	Qualcomm's Interest in Intel Could Spell Troub...	Qualcomm is rumored to be interested in acquir...	The Motley Fool	2024-09-30T09:10:00Z

Figure 3. Raw news data

30,000 english news articles were collected on 3 stocks from 10 major financial news publishers: *Zacks Investment Research, Investing.com, MarketWatch, Invezz, The Motley Fool, Seeking Alpha, Benzinga, Quartz, PennyStocks, and GlobeNewswire Inc.* On average, 8.7 news per day are collected for all stocks.

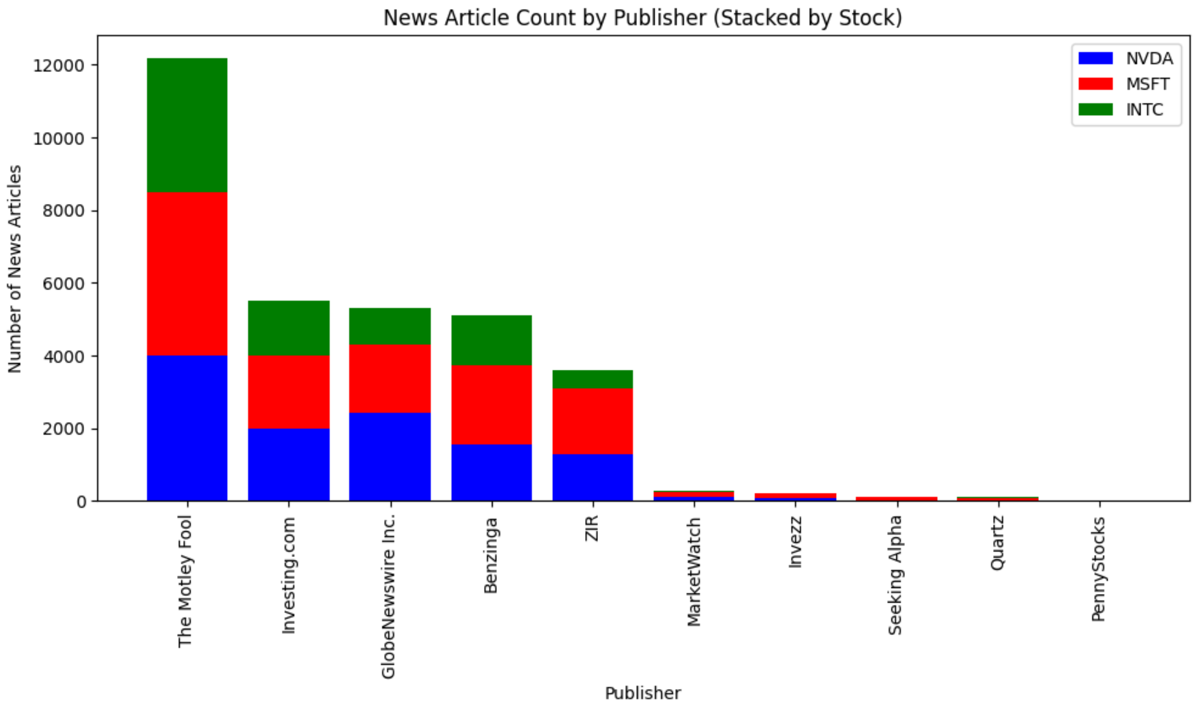


Figure 4. News count by publishers.

The *Financial PhraseBank* is a specialized dataset for financial sentiment analysis, containing 4,845 English news from LexisNexis financial news. Each sentence of the news is carefully annotated by 16 finance and business experts. We selected this dataset to enhance our pretrained NLP model *FinBERT*'s accuracy in sentiment scoring.

0	neutral	Technopolis plans to develop in stages an area...
1	negative	The international electronic industry company ...
2	positive	With the new production plant the company woul...
3	positive	According to the company 's updated strategy f...
4	positive	FINANCING OF ASPOCOMP 'S GROWTH Aspocomp is ag...
...	...	...
4840	negative	LONDON MarketWatch -- Share prices ended lower...
4841	neutral	Rinkuskiai 's beer sales fell by 6.5 per cent ...
4842	negative	Operating profit fell to EUR 35.4 mn from EUR ...
4843	negative	Net sales of the Paper segment decreased to EU...
4844	negative	Sales in Finland decreased by 10.5 % in Januar...

4845 rows × 2 columns

Figure 5. Financial PhraseBank dataset

3.2.2 Data processing

After collecting the raw news data, we applied standard NLP preprocessing: converting dates to datetime format, removing duplicate and missing entries, eliminating redundant text and titles, and filtering out stopwords. We also merged title and content together for the sentiment scoring process.

While news data was gathered daily, stock data was discontinuous due to market closures on weekends and holidays. To address this, we merged news from closed-market days with the nearest open-market day. For example, news from Saturday and Sunday was merged with Monday's data.

3.3 Training *FinBERT*

FinBERT is a specialized Natural Language Processing (NLP) model designed for financial sentiment analysis, built upon the *BERT* (Bidirectional Encoder Representations from Transformers) architecture.

While *FinBERT* performs exceptionally well in scoring sentiment for financial news and is among the top pretrained NLP models for this task, there is still room for improvement. We used the *Financial PhraseBank* to fine-tune the model, splitting the dataset into 70% for training, 20% for testing, and 10% for validation. The fine-tuned model demonstrated significant improvements in key performance metrics.

Models	**Precision**	**Recall**	**F1-Score**	**Accuracy**
FinBERT	0.930	0.931	0.930	0.932
Fine-tuned *FinBERT*	0.969	0.973	0.970	**0.974**

Table 1. Fine-tuned *FinBERT* metrics

3.4 Sentiment Scoring and Feature Extraction

The processed news data was input into the fine-tuned FinBERT model to obtain sentiment scores, with each news item categorized as Positive, Neutral, or Negative, based on a score from 0 to 1. Each news item is assigned the category with the highest score among these three.

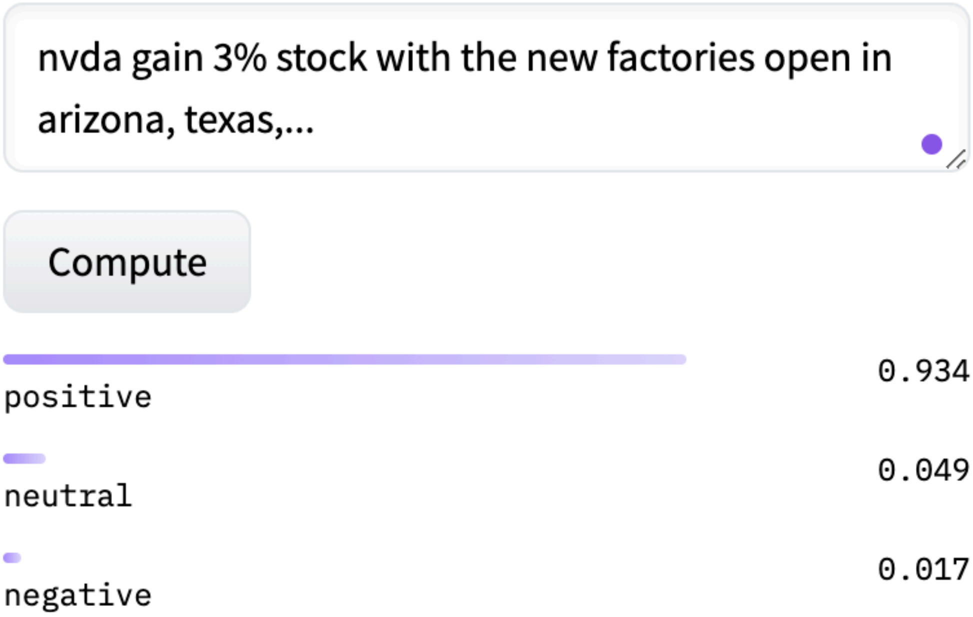


Figure 6. Example usage of FinBERT

For this project, we adjusted sentiment values to a range of -1 to 1, where values close to -1 represent negative sentiment, and values close to 1 indicate positive sentiment. If FinBERT predicts a Positive score, we retain it as-is; for Negative, we take the negative of the score; and for Neutral, we set it to 0.

To calculate daily sentiment, we applied a weighted average, emphasizing news between 9:30 am and 4:30 pm—trading hours when investors are more likely to make buy or sell decisions,also for day before market open again (Sunday, end of Holidays), we used higher weight due to it allows investors to react to news and events that occurred during the break, leading to more informed trading decisions.

Time	**Weight**
12:00 AM to 9:29 AM	0.8
9:29 AM to 4:30 PM	1.2
4:30 PM to 11:59 PM	0.8
Regular days	1.0
Day before market open again (Sunday, end of Holidays)	1.2

Table 2. Weight Calculation of sentiment

Next, we calculated the daily sentiment and merged it with the stock data, aligning the dates accordingly. We removed the *Open, Close, High,* and *Low* columns, as they were not relevant to our project. Instead, we used the *Adjusted Close*, which is better suited for long-term analysis as it accounts for corporate actions like dividends and stock splits, offering a more accurate measure of investor returns.

	Date	Adj Close	Volume	sentiment
0	2021-04-27	15.349164	164572000	0.380556
1	2021-04-28	15.244386	209416000	-0.080159
2	2021-04-29	15.292284	173196000	0.648261
3	2021-04-30	14.977703	201912000	0.721436
4	2021-05-03	14.805319	203912000	0.200206
...	...	...	...	...
865	2024-10-03	122.849998	277118000	0.320207
866	2024-10-04	124.919998	243678100	0.398219
867	2024-10-07	127.720001	346250200	0.282160
868	2024-10-08	132.889999	285722500	-0.232941
869	2024-10-09	132.649994	246191600	0.153954

Figure 7. Processed Data

3.5 Analysis and feature engineering

In this section, we combined sentiment with stock data to analyze correlations and uncover patterns. This approach allows us to evaluate the impact of sentiment on stock movements and improve forecasting models by incorporating sentiment as a predictive feature for price changes.

Firstly, we calculated the correlation between each feature, then plotted a heatmap to visually identify which features have the strongest relationships with sentiment.

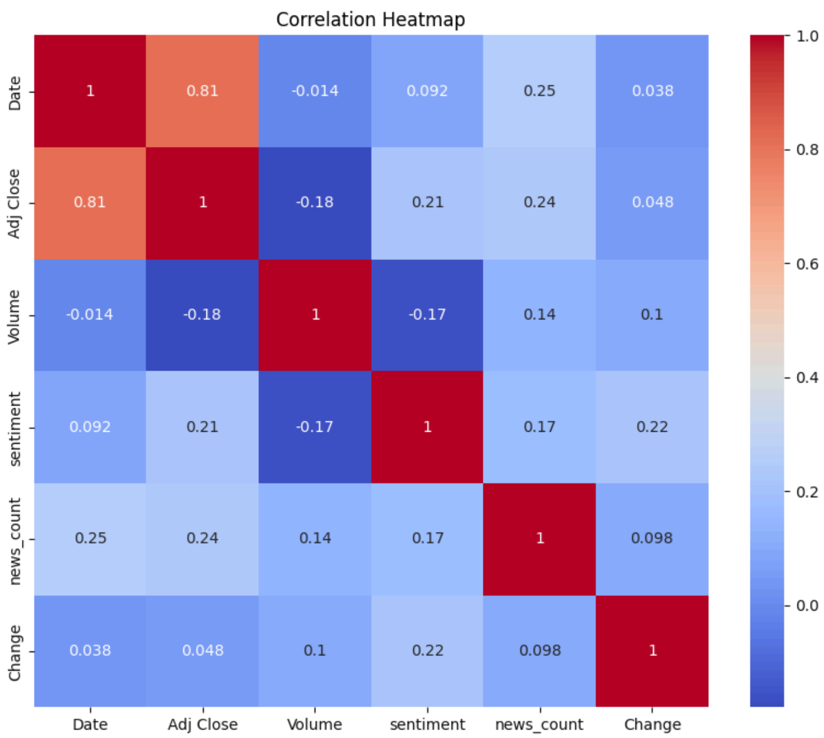


Figure 8. Correlation heatmap of NVDA's feature

For all three stocks, the correlation patterns are quite similar. The heatmap shown above represents the correlations for NVDA, indicating that the top 3 features most relevant to sentiment:

Stocks\Features	Change	Adj Close	News Count
NVDA	0.2154	0.2131	0.1744
MSFT	0.3246	0.3219	0.2732
INTC	0.2308	0.2294	0.2432

Table 3. Correlation of top3 feature most relevant to sentiment

These value reveal that sentiment alone may not be a strong predictor and should be combined with other indicators for more robust analysis

3.5.1 Change of stock price and sentiment

Change in MSFT stock price showed the highest correlation with sentiment scores, we initially focused on plotting these two variables together and coloured them yellow if they both increase or decrease, and blue if not.

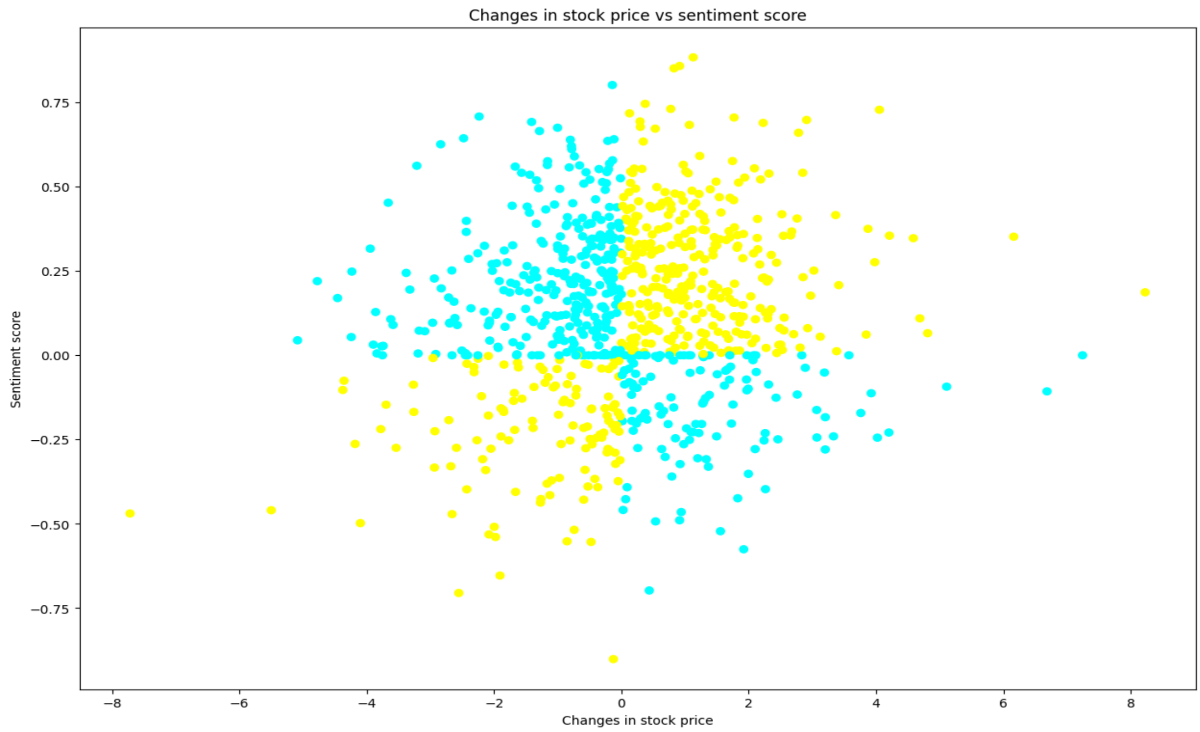


Figure 9. Change in stock price vs sentiment score of MSFT

However, the pattern is not clearly visible to the naked eye. This led us to create a new feature by multiplying the stock price change with the sentiment score, essentially capturing when both metrics move in the same direction.

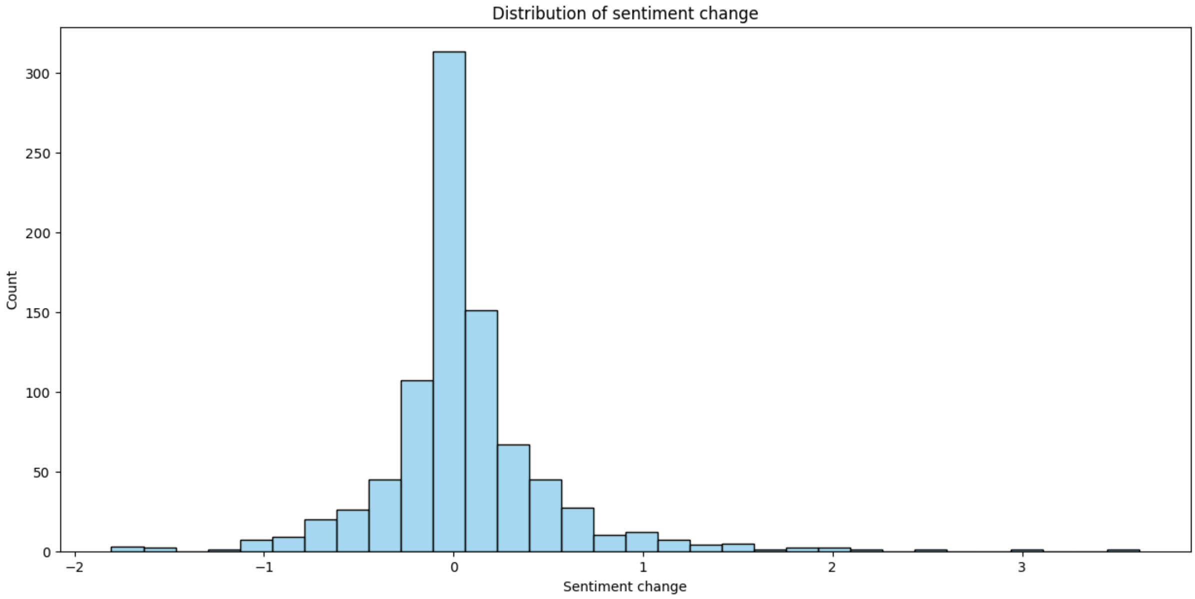


Figure 10. Distribution of Sentiment x Change of INTC

The histogram of this new 'sentiment-change' tends to be distributed to the right, all three stocks have more than 51% of the instances positive. It indicates that more often than not, the sentiment and price changes moved in the same direction. This slight majority, while not overwhelming, supports the effectiveness of our sentiment model and aligns with basic market theory - that positive sentiment tends to drive prices up, while negative sentiment tends to drive them down.

Then, we draw a box plot that displays the distribution price changes across different sentiment categories (Very Positive, Positive, Neutral, Negative, Very Negative). Within the report, we show the plot of NVDA, which is considered our most unpredictable stock.

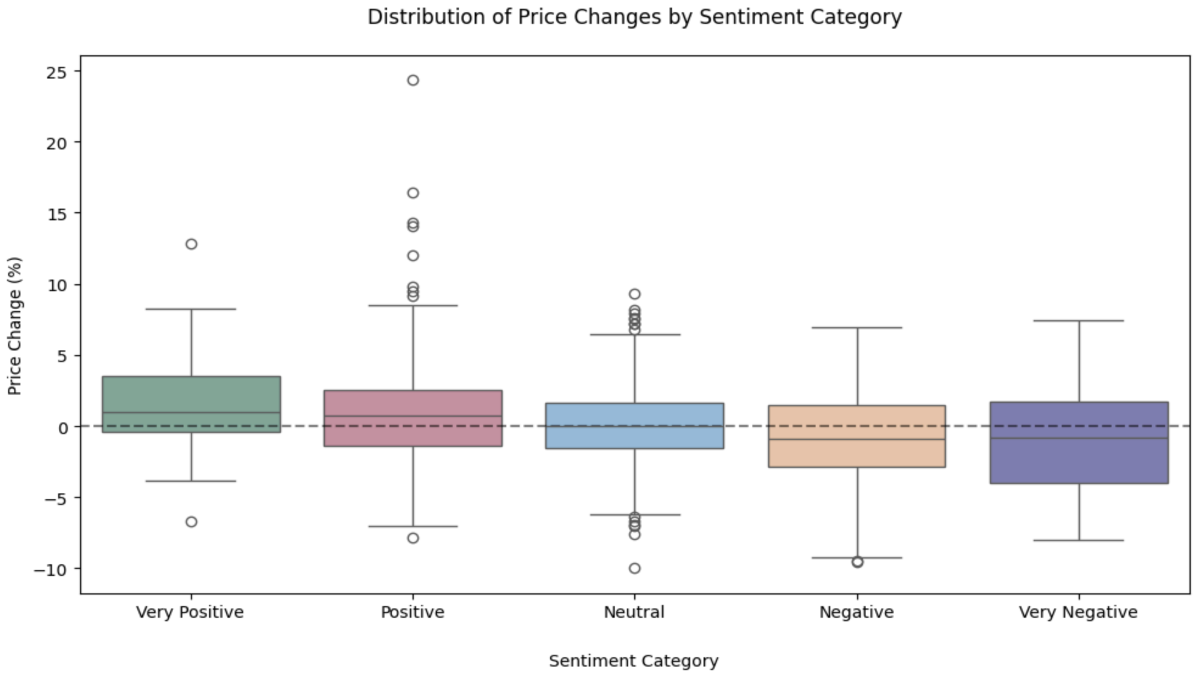


Figure 11. Distribution of price change by sentiment category

In this visualization, the plot shows a clear relationship between the sentiment category and the distribution of price changes. **Positive sentiment** categories (Very Positive and Positive) are associated with **higher median** price changes, while **negative sentiment** categories (Negative and Very Negative) are associated with **lower median** price changes.

The finding validates our sentiment analysis approach, as it captures the fundamental relationship between market news and price movements.

3.5.2 Stock price and sentiment

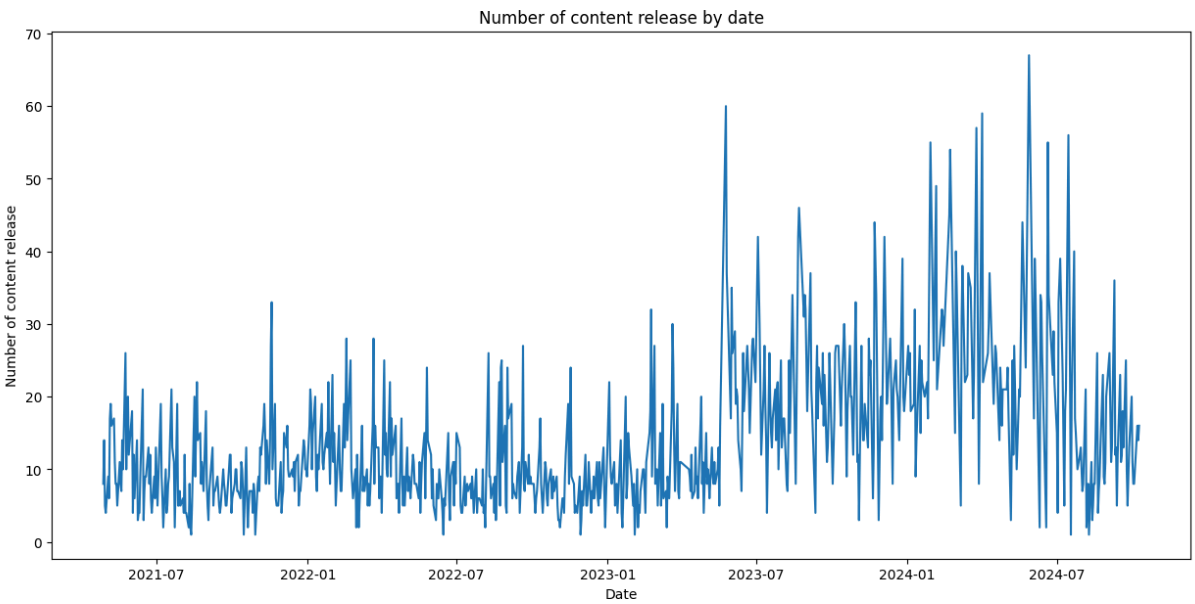


Figure 12. Number of NVDA news release by date

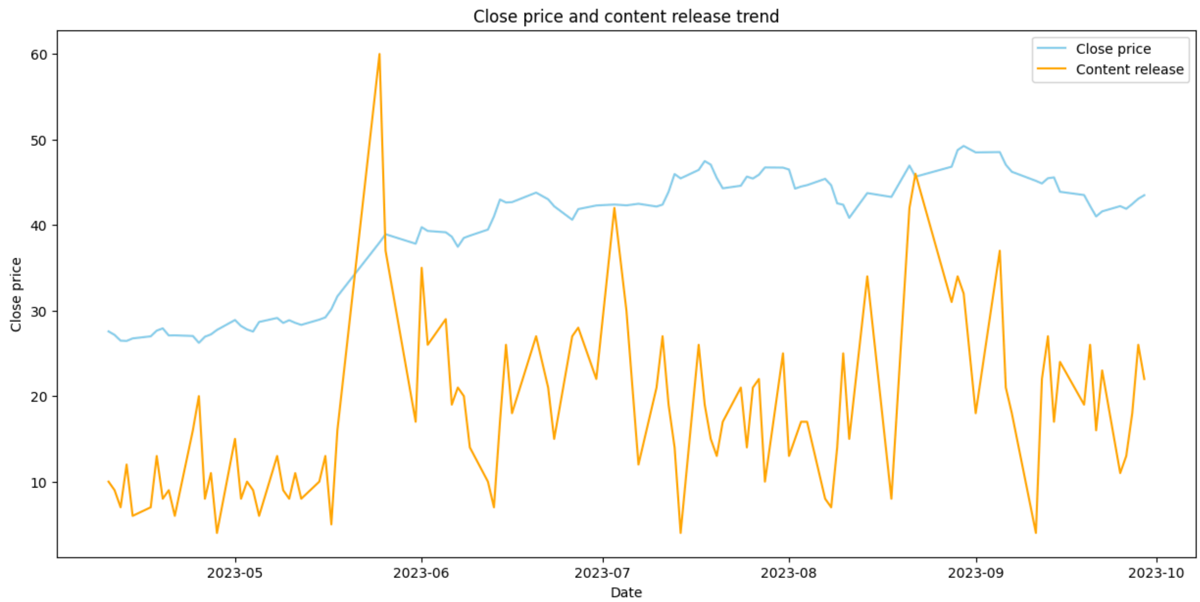


Figure 13. Close price and content release trend

Looking at the news release chart of NVDA, we see a clear change between the 2 time periods before and after June 2023. The reason for this change could be the spike in the middle of May. It not only shows the close relationship between number of news and price but also reveals that the market has undergone certain changes. Therefore, we plot the candlestick chart with sentiment score on two periods to ensure consistency.

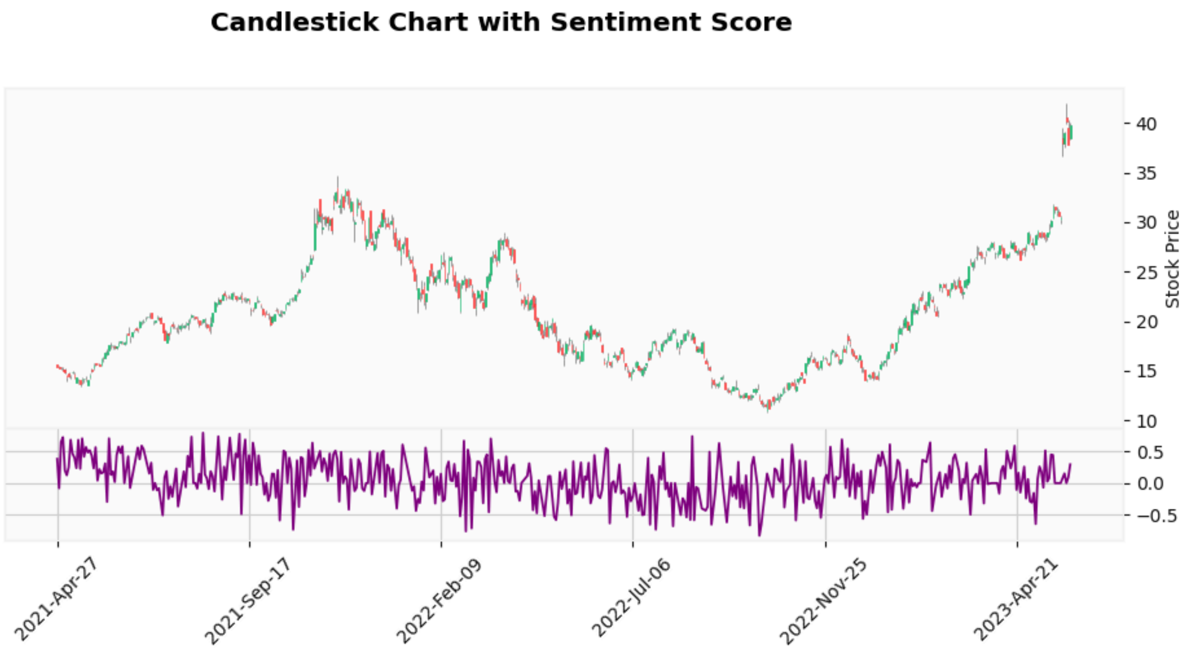


Figure 14. Candlestick chart with sentiment score before June 2023

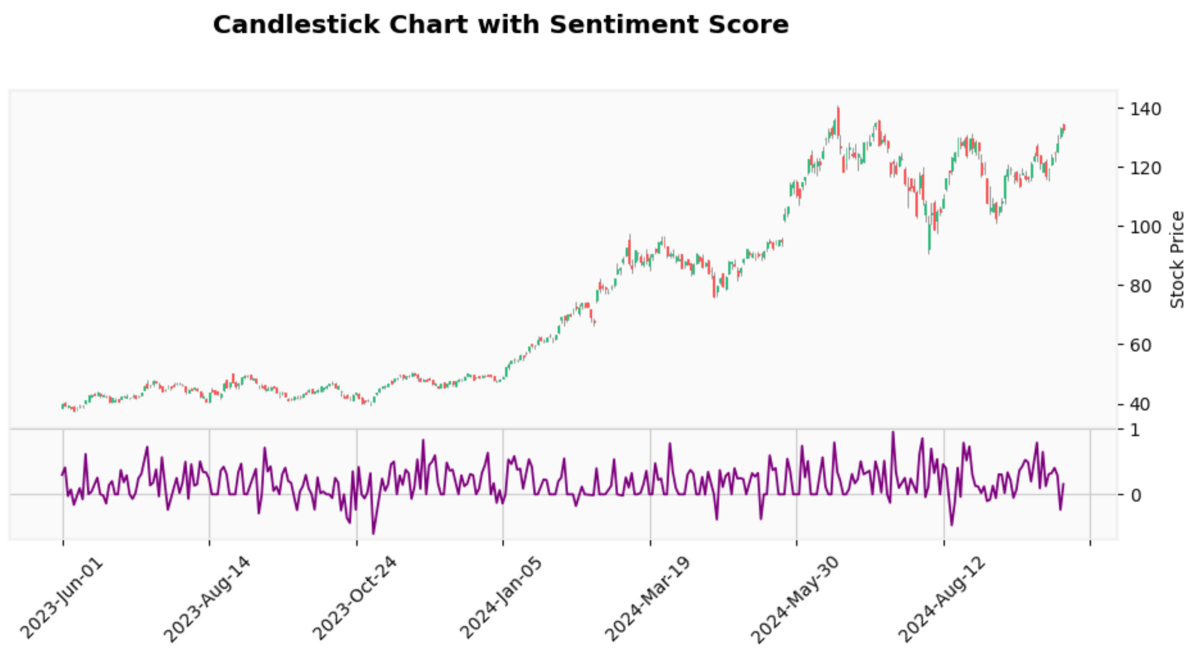


Figure 15. Candlestick chart with sentiment score after June 2023

This plot combines a candlestick chart for the stock price with an overlay of sentiment scores over time. Observing two charts, we see a clear difference, while the first chart shows constant fluctuations without following a trend, the second chart immediately shows us an upward trend in the stock price suggesting that we should focus on the second one.

The chart shows clear uptrends and downtrends, with some periods of volatility, while the sentiment score line (in purple) shows frequent oscillations but stays mostly between 0 and 1 for the second. Focusing on possible correlation, we see that periods of increased sentiment activity (such as spikes) roughly align with certain price movements. We do the same to two remaining stocks then it all suggests that sentiment spikes could contribute to volatility or coincide with significant market events.

3.5.3 Moving average

In this section, we want to smooth noise and identify longer-term trends and convergence or divergence patterns, which might reveal sentiment's predictive qualities for stock movement. We consider using the moving average and the volatility but the correlation of the **moving average** and sentiment return a better signal: the mean value is approximately 0.20 for moving average and -0.15 for volatility.

The analysis shows that sentiment data and stock prices respond differently. Sentiment analysis benefits from both short and long-term moving averages, likely due to its volatile nature in reflecting market psychology.

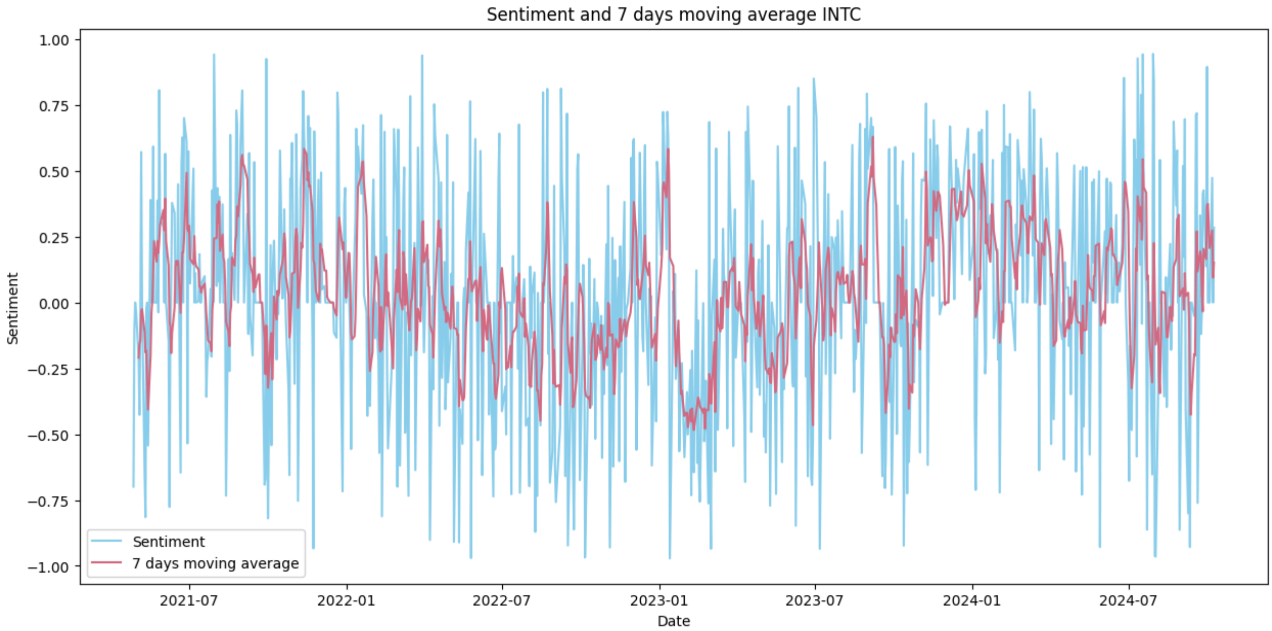


Figure 16. 7 days moving average and sentiment

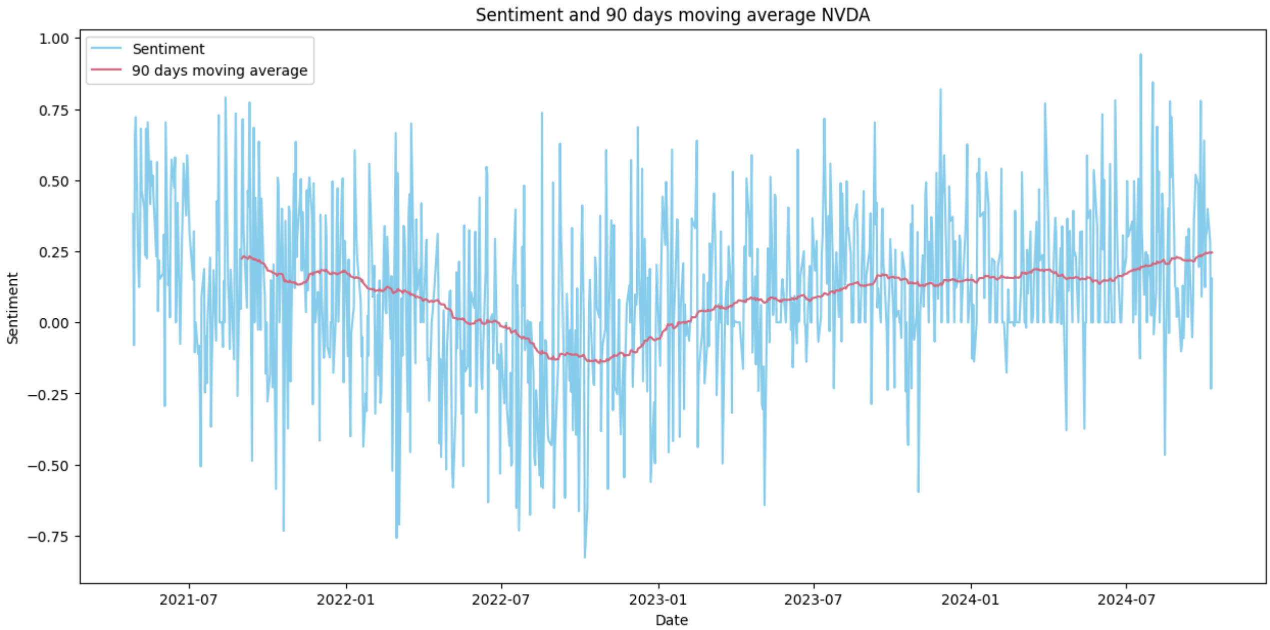


Figure 17. 90 days moving average and sentiment

In contrast, stock prices align great only with shorter 7-day moving averages, as they require more responsive indicators to capture market movements effectively.

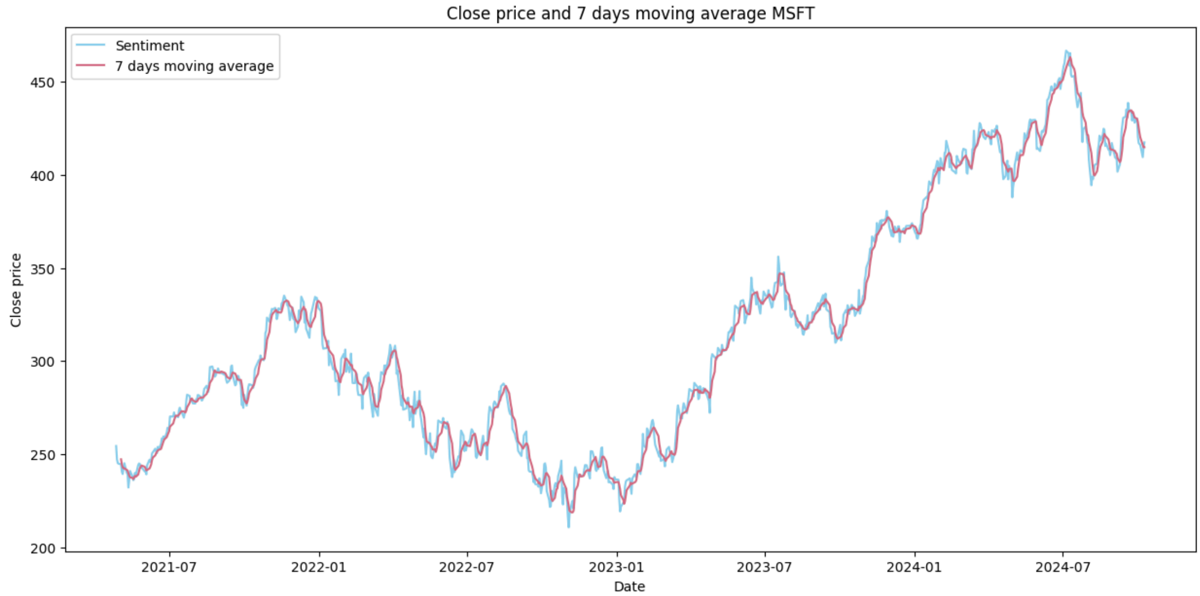


Figure 17. 7 days moving average and close price

In conclusion, after analysis of the correlations and visualizations, we determined that **Adjusted close, Change of stock, News count, 7-day Moving average and Sentiment score** are the most relevant features for predicting stock price movements. The sentiment analysis revealed a moderate correlation with short-term price changes, particularly in the context of shorter moving averages. The 7-day moving average proved effective in capturing recent trends, aligning well with both sentiment and price shifts, while adjusted close provides the core stock price information. Based on these findings, we selected these three features to feed into the predictive model, focusing on capturing both recent price momentum and sentiment-driven market shifts.

	Date	Adj Close	sentiment	news_count	Change	moving_avg_7
0	2021-04-27	15.349164	0.380556	8.0	-0.100836	15.234621
1	2021-04-28	15.244386	-0.080159	14.0	-0.104778	15.222504
2	2021-04-29	15.292284	0.648261	5.0	0.047898	15.244386
3	2021-04-30	14.977703	0.721436	4.0	-0.314581	15.194349
4	2021-05-03	14.805319	0.200206	9.0	-0.172384	15.192425
...	...	...	...	...	...	...
865	2024-10-03	122.849998	0.320207	8.0	4.000000	121.298572
866	2024-10-04	124.919998	0.398219	10.0	2.070000	121.500000
867	2024-10-07	127.720001	0.282160	16.0	2.800003	122.025714
868	2024-10-08	132.889999	-0.232941	14.0	5.169998	123.667143
869	2024-10-09	132.649994	0.153954	16.0	-0.240005	125.268570

Figure 18. Final Processed Data

3.6 Modeling

3.6.1 LSTM

LSTM (Long Short-Term Memory) networks are a type of RNN (Recurrent Neural Network) designed to handle sequential data and overcome RNNs' limitations, especially the vanishing gradient problem, which hinders learning long-term dependencies. LSTMs achieve this with memory cells and three gates—input, forget, and output—that manage information flow, enabling them to retain or discard data as needed.

In stock price prediction, LSTMs are effective when using features like past prices and sentiment scores from news.

3.6.2 Data preparation

We use 85% of our data for training and use the remaining 15% for testing to evaluate the performance of our model.

We divided the features into 2 following categories:

- Lagged features: **Adjusted close, Change of stock, 7-day Moving average**

- For each *t* day , the input will include stock prices from the previous *7* days
- Our sequence length is 7, then to predict the price at day *t*, the model would include stock prices from days *t−7* to *t−1*.

- Real-time features: Sentiment, News Count

- Sentiment score and news count are used **in real time**—meaning they are specific to the prediction day, *t*.
- Each day's sentiment and news count values are appended to the lagged stock price sequence for that day's prediction.

3.6.3 Model Design

The LSTM will process the lagged stock price sequence to learn trends and patterns over time, while the added sentiment and news count inform the model of the current day's market context.

This combined approach enables the model to predict day *t* based on both historical stock trends and the immediate sentiment.

Example Input Structure for Day *t*:

- X_train (for day *t*): [*Adj_close_t-10, Adj_close_t-9, ..., moving_average_7_t-1, sentiment_t, news_count_t*]
- Y_train (target): *Adj_close_t* (the stock price on prediction day)

4. Results and Finding

In this section we evaluate the model we built with 3 metrics to access the performance: **MAE, MAPE, Accuracy**

We evaluate the impact of additional features on the performance of the LSTM model by comparing two scenarios: one where the model uses only the historical stock data as the features for prediction, and another where the LSTM-FinBERT model incorporates sentiment features, including the **news count, and sentiment score.**

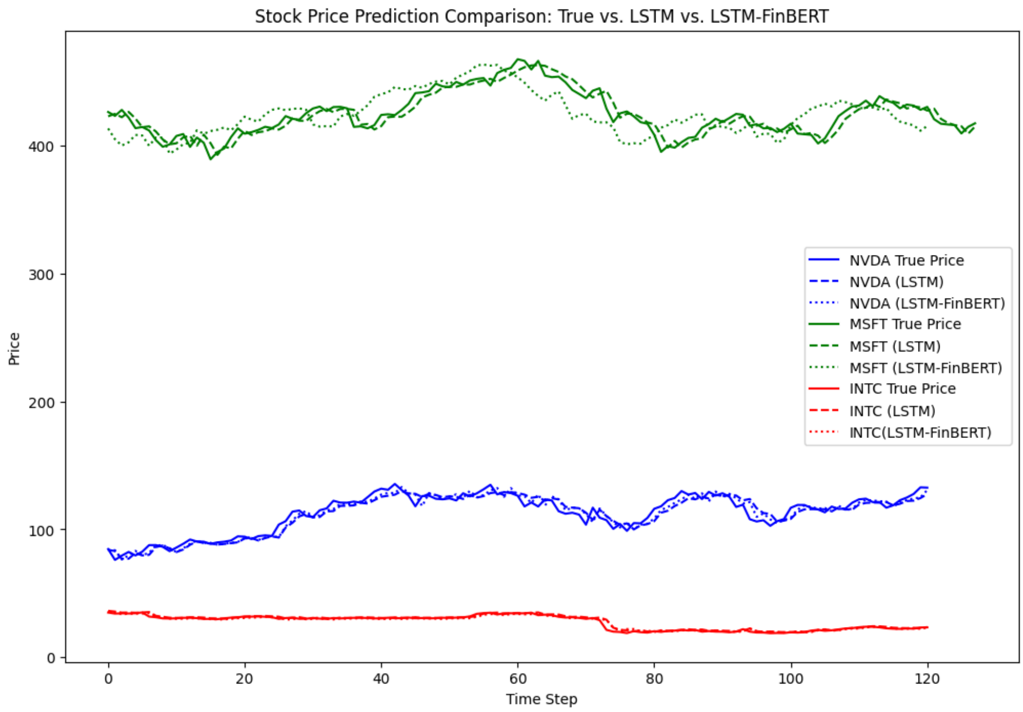


Figure 19. Comparison stock price on True vs. LSTM vs. LSTM-FinBERT

From the plot of Fig 19. we can see the predictions of FinBERT-LSTM are closer to the actual price than the LSTM model.

Model	**MAE**	**MAPE**	**ACCURACY**
NVDA (LSTM)	3.98	0.035	0.964
NVDA (LSTM-FinBERT)	3.34	0.018	0.971
MSFT (LSTM)	4.40	0.010	0.989
MSFT (LSTM-FinBERT)	4.01	0.007	0.992
INTC (LSTM)	0.732	0.027	0.972
INTC (LSTM-FinBERT)	0.633	0.018	0.981

Table 4. Model evaluation

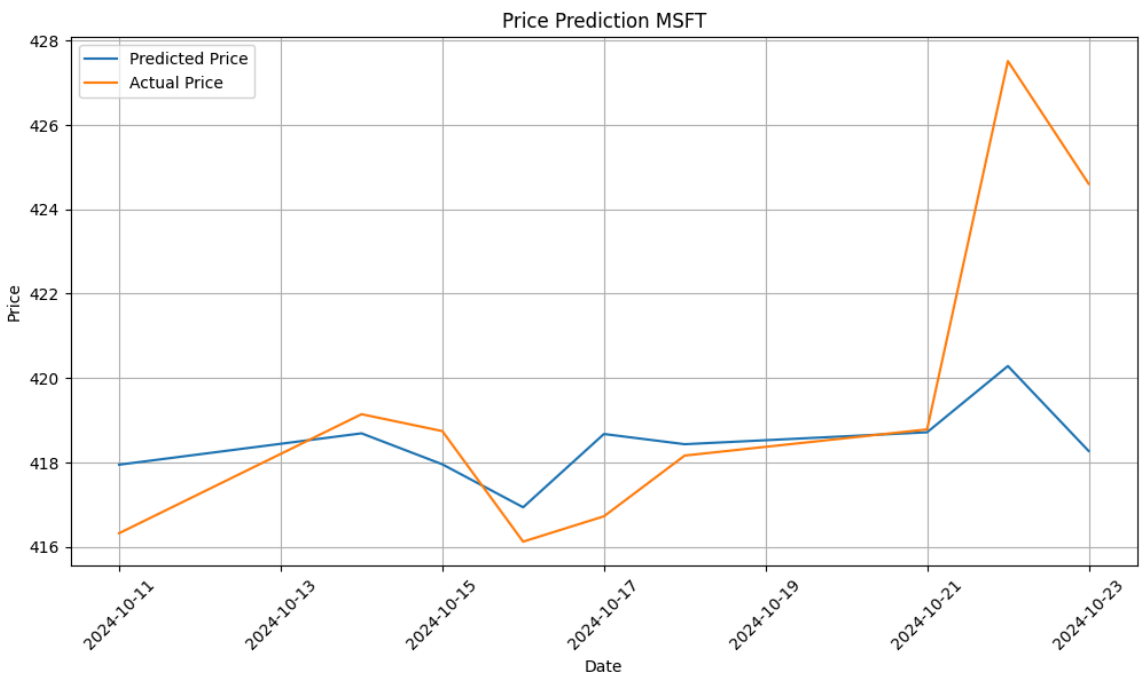
We observe that the LSTM-FinBERT model improves its accuracy from 0.3% to 0.9% compared to the original LSTM, which utilized only a single feature. This significant increase demonstrates a strong correlation between the selected features and the predicted price in the test set.

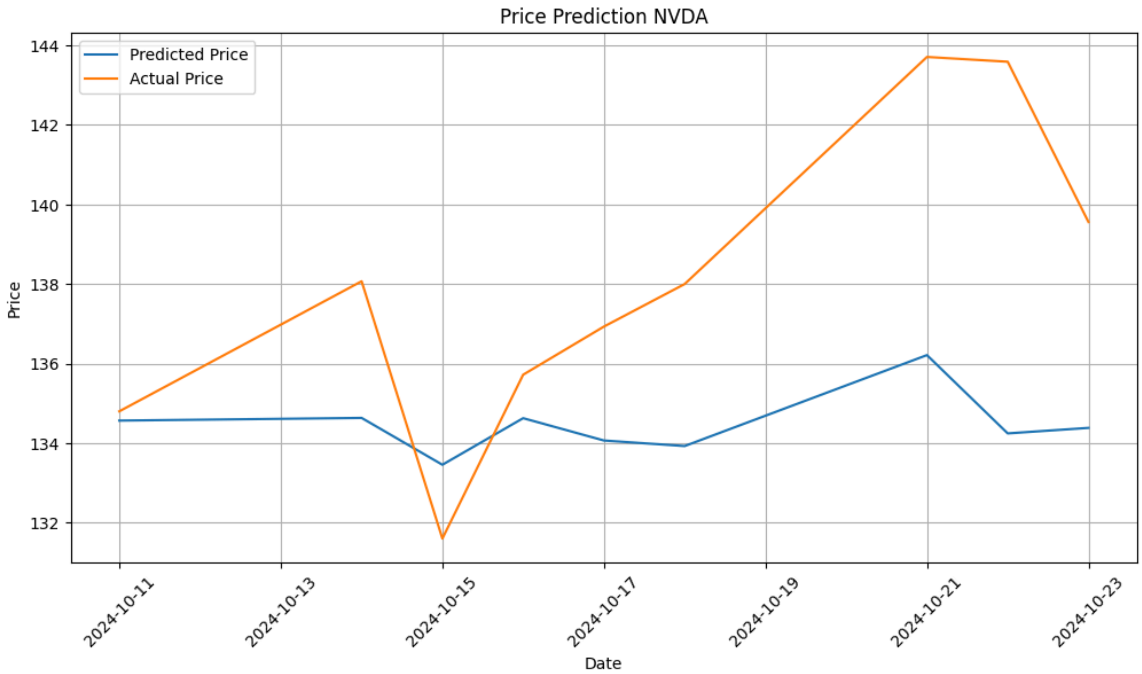
The results are obtained after running multiple trials on the best performing architecture. We ran 100 trials on each model to mitigate randomness.

5. Forecasting Future Price

In this section, we utilized the saved LSTM-FinBERT model that has been trained on historical data to predict future outcomes. The input data for the model will consist solely of the news from the day being predicted, along with the number of news articles. Price-related features will be derived from past data, serving as lagged features.

The unseen data will be processed in the same way as the historical data, ensuring that the processed dataset includes the sentiment for that day and the corresponding news count.





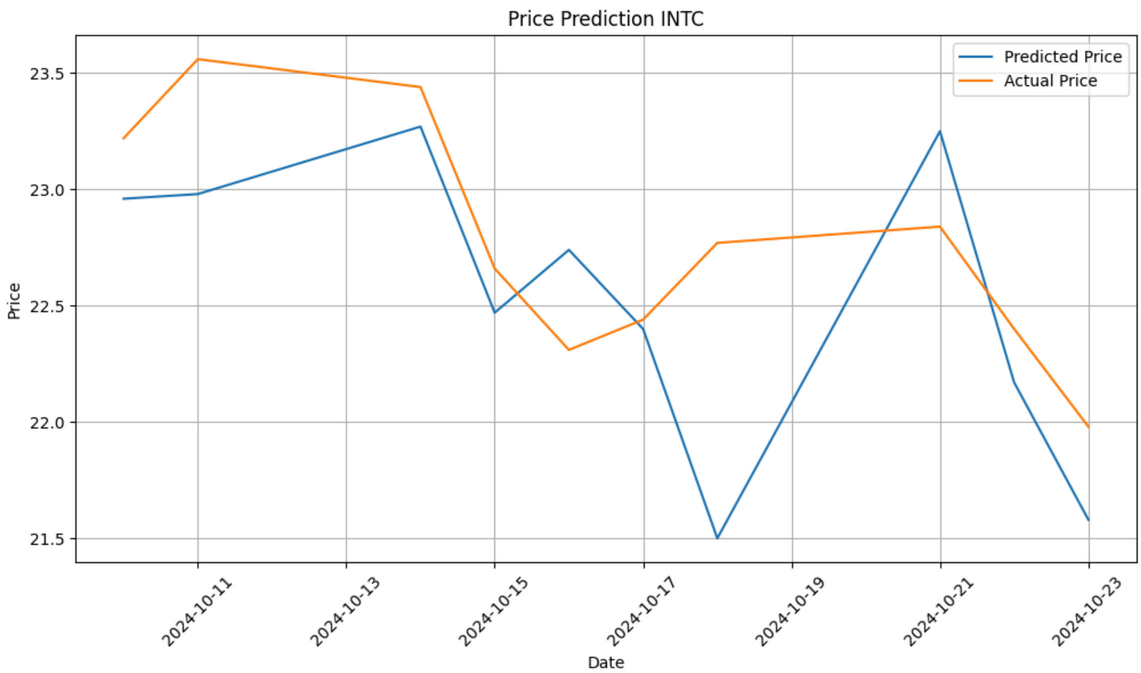


Figure 20. Stock prediction for 3 stocks

We evaluated the prediction with 2 metrics to access the performance: **MAE, Direction Accuracy:**

Stocks	**MAE**	**Direction accuracy**
NVDA	38.03	62.5%

MSFT	12.93	75%
INTC	2.39	50%

Table 5. Model Evaluation

We can observe that while the overall direction of the stock prediction is reasonably accurate, the predicted price itself is not closely aligned with the actual results, especially for harder stocks to predict like NVDA.

6. Discussion

The LSTM-FinBERT model demonstrates commendable performance for a challenging task, such as predicting unseen stock data. By incorporating sentiment from news articles and various price-related factors, the model effectively aids investors in making informed decisions regarding potential upward or downward movements in stocks. This is particularly valuable in a market environment where stock prices are significantly influenced by news and rumors.

However, it is important to recognize that sentiment and news alone cannot fully capture the complexities of stock price movements. There are numerous other factors that can impact stock performance, including insider trading information, inflation rates, employment statistics, and broader economic indicators. These elements play a crucial role in shaping market dynamics and can lead to price fluctuations that are not reflected in sentiment analysis.

In future work, several enhancements could be explored to improve stock price prediction using the LSTM-FinBERT model. Implementing ensemble learning techniques could combine predictions from multiple models, such as traditional machine learning algorithms and advanced neural network architectures like Transformers, to enhance accuracy and robustness. Additionally, expanding the feature set through technical indicators, fundamental data, and broader market sentiment from social media could provide deeper insights into price movements.

Furthermore, integrating real-time data for continuous updates on sentiment and price-related features would make the model more relevant for traders in fast-paced environments. Longitudinal studies tracking model performance across different market conditions could also refine its adaptability, ensuring that investors have comprehensive tools for navigating the complexities of financial markets.

7. Bibliography

[1] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[2] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," in Proc. Applied Natural Language Processing Conference, pp. 2-12, 2019.
P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, "Good debt or bad debt: Detecting semantic orientations in economic texts," Journal of the Association for Information Science and Technology, vol. 65, no. 4, pp. 782-796, 2014.

[3] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review," Applied Soft Computing, vol. 90, pp. 106181, 2020.

[4] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," The Journal of Finance, vol. 25, no. 2, pp. 383-417, 1970.

[5] M. Baker and J. Wurgler, "Investor Sentiment and the Cross-Section of Stock Returns," The Journal of Finance, vol. 61, no. 4, pp. 1645-1680, 2006.

[6] P. C. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," The Journal of Finance, vol. 62, no. 3, pp. 1139-1168, 2007.

[7] Y. Li, P. C. Tetlock, and Y. Zhang, "The Real-Time Price Impact of Corporate News: Evidence from Natural Language Processing," Review of Financial Studies, vol. 34, no. 8, pp. 3980-4021, 2021.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. North American Chapter of the Association for Computational Linguistics, pp. 4171-4186, 2019.

[9] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep Learning for Event-Driven Stock Prediction," in Proc. International Joint Conference on Artificial Intelligence, pp. 2327-2333, 2015.

[10] Polygon.io API Documentation (2023). "Financial Market Data API"1
.
[11] Yahoo Finance API Documentation (2023). "Market Data and Company Information"

[12] L. Bellintani, "The Effect of News on Intra-Day Stock Prices & their Volatility," Student Economic Review, vol. XXXIII, 2019. [Online]. Available: https://www.tcd.ie/Economics/assets/pdf/SER/2019/11stock_prices.pdf.

[13] M. Nesvisky, "Which News Moves Stock Prices?," NBER Digest, Jun. 2013. [Online]. Available: https://www.nber.org/digest/jun13/which-news-moves-stock-prices.