Integrating Sentiment Analysis to Enhance Stock Prediction and Investment Strategies at Blackstone

1. *Executive Summary*

In the modern age of investing, increasing complexity arises in predicting market movements due to volatility driven by non-traditional market forces, including social media trends, and real-time news events. In this pilot, we will be integrating sentiment analysis into Blackstone's investment strategies by utilising real-time data from financial news and social media platforms to create a comprehensive dashboard. We developed a robust dashboard that incorporates sentiment data from multiple sources, including New York Times API (in-depth financial reporting) and Reddit API (discussions from retail investor communities such as r/Investing), to enhance prediction of stock performance and optimise investment strategising.

By utilising this real-time data, we aim to combat the stochastic nature of the market by factoring in individual opinion through sentiment. The platform is user-friendly and easy to integrate with Blackstone's existing analytics infrastructure, enabling seamless use of sentiment insights alongside traditional financial data. This solution will allow Blackstone to improve investment decision-making, optimise asset allocation, mitigate potential risks and capitalise on market momentum while maintaining a competitive edge.

2. *Statement of the Problem*

Blackstone faces increasing complexity in predicting market movements due to the volatility driven by non-traditional market forces, including social media trends, and real-time news events. The rise of platforms like Reddit and Twitter has amplified market noise, making it harder to discern valuable insights from speculation. Traditional data sources alone are no longer sufficient for maintaining an edge.

Given Blackstone's scale and focus on generating superior returns for institutional investors, leveraging innovative data analytics tools to stay ahead of market trends is essential. In this context, our team proposes to implement sentiment analysis, using real-time news and social media data, to better predict stock performance and deliver actionable insights that can complement your existing investment frameworks. This solution will allow Blackstone to improve investment decision-making, optimise asset allocation, mitigate potential risks and capitalise on market momentum.

3. *Research Questions*

Research Question #1: Compared to sentiment derived from news articles, would sentiment derived from social media platforms like Reddit lead to more accurate predictions of stock performance?
Null Hypothesis (H0): There is no difference in the accuracy of stock performance predictions between sentiment derived from news articles and sentiment derived from social media platforms like Reddit.
Alternative Hypothesis (H1): Sentiment derived from social media platforms like Reddit leads to more accurate predictions of stock performance than sentiment derived from news articles.
Metrics: The correlation coefficient between sentiment scores and actual short-term stock price movements. The model's mean absolute percentage error for both sentiment sources will also be used to assess prediction accuracy.
Research Question #2: Compared to neutral or stable sentiment trends, would extreme positive or negative sentiment shifts in news and social media predict higher stock price volatility?

Null Hypothesis (H0): Extreme positive or negative sentiment shifts in news and social media do not predict higher stock price volatility compared to neutral or stable sentiment trends.
Alternative Hypothesis (H1): Extreme positive or negative sentiment shifts in news and social media predict higher stock price volatility compared to neutral or stable sentiment trends.
Metrics: Stock price volatility will be measured using the standard deviation of stock prices over a defined time window, as well as the volatility index (VIX) where applicable.

4. Data Exploration

Daily News Sentiment Index:

Extracted from news articles, the **news sentiment index** dataset contains sentiment scores reflecting the positive or negative tone of media coverage, providing insights into the general market sentiment and its impact on stock prices. It comprises daily news scores since January 1, 1980. Generally, a higher score indicates a more positive public sentiment. This dataset has a total number of **16150** rows and **2** columns.

Combined_nasdaq_yf_data_2023:

This file contains the data of **2023 Nasdaq stock price details** which provides more specific data about the stocks in a certain time period that we will use to analyse combined with the sentiment index. By doing that we believe it can generate deeper understanding between the market sentiment and the actual fluctuations of stock prices. This dataset has a total number of **1048576** rows and **23** columns.

News API & NYTimes API:

The News API and NYTimes API provide real-time access to a vast collection of news articles from various sources. By using these APIs, we can continuously gather and update the news sentiment data in real-time. This data will complement the historical sentiment index, offering a more current view of media coverage and market sentiment. By scraping with these APIs, it allows us to query articles by date, keyword, or sentiment, enabling us to align media narratives with stock price movements and market conditions.

Reddit API:

The Reddit API allows access to user-generated content from one of the largest social media platforms. By scraping data and analysing discussions on subreddits related to finance, stocks, and market trends, we can capture public sentiment from a diverse audience of retail investors and market enthusiasts. This dataset provides a unique perspective on market sentiment, complementing traditional media by reflecting grassroots sentiment and market rumours, which often impact stock prices.
utility index (VIX) where applicable.

5. Data Collection and Integration (Samaa Nadkarni)

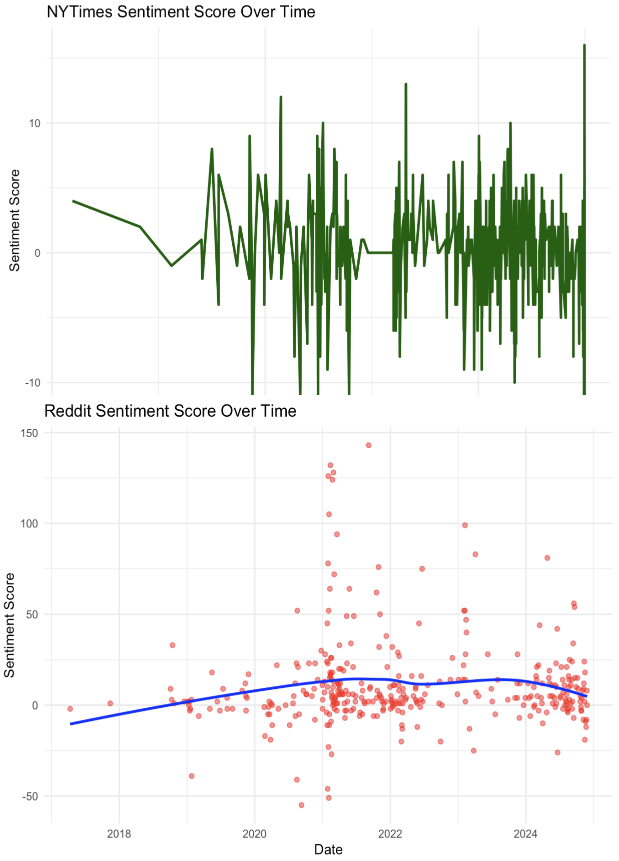
The integration of API connections to the New York Times and Reddit marks a significant milestone in establishing a comprehensive data acquisition framework. These connections enable the continuous collection of both historical and real-time data from a reputable news outlet and an influential social media platform, ensuring a robust foundation for financial sentiment analysis. Data gathering operations are strategically structured to cover a wide spectrum of relevant content, with curated queries targeting key financial topics such as stock market fluctuations, sector-specific trends and investor sentiment.

For the New York Times, the script leverages their API to fetch articles related to stock market fluctuations across the energy, healthcare, and technology sectors using a list of search queries. This list can be modified depending on the client's needs to do an API fetch on a different subject matter. It systematically extracts metadata such as headlines, abstracts, URLs, and publication dates using dynamic query generation and iterative API calls. Rate-limiting compliance is managed through delays between requests, while error handling ensures resilience against empty or failed responses. The current free tier API access allows retrieval of article abstracts, not the entire article body. While the abstracts are valuable, premium access could enhance capabilities, reduce latency and support uninterrupted data flow. The decision to upgrade to premium access will depend on pilot results evaluating performance and scalability.

Similarly, the Reddit API integration focuses on subreddits such as "investing" and "wallstreetbets," which are rich in user-generated discussions about financial markets. The script authenticates using OAuth and retrieves metadata for the top posts, including titles, authors, timestamps, upvotes and comments. Advanced filters ensure relevance by isolating financial content while excluding non-essential data. A structured iteration through subreddits enables the aggregation of insights into investor sentiment and market discussions, while real-time streaming optimisations reduce latency and improve data flow.

Initial data cleaning protocols address issues such as missing values, duplicates and inconsistent formatting, ensuring compatibility and readiness for downstream analysis. Advanced preprocessing scripts standardise diverse data formats, while monitoring systems ensure the integrity and quality of incoming data. The pipeline is designed to handle vast volumes of unstructured data from Reddit while maintaining compliance with API guidelines, user privacy and ethical standards.

Combined, these efforts form a robust and scalable data acquisition system capable of supporting real-time sentiment analysis and predictive modeling. These efforts lay a solid foundation for predictive modeling and actionable insights, ensuring that Blackstone's investment strategies are informed by the most accurate and timely data available.

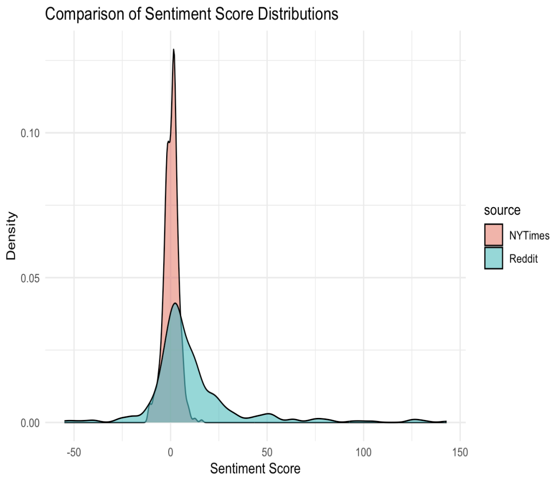


6. *Sentiment Analysis (Leqi Han)*

Sentiment analysis was conducted using the AFINN lexicon, a tool specifically designed for scoring the emotional tone of textual data. The analysis leveraged two primary datasets: financial news articles obtained via the New York Times API and user-generated discussions from Reddit subreddits.. These sentiment sources were integrated with historical Nasdaq stock price data to investigate the impact of sentiment on stock performance and market volatility.

The process began with data cleaning and preparation. Text data from both sources were tokenized into individual words, and common preprocessing steps such as removing stopwords, punctuation, and non-informative tokens were applied. Sentiment scores for each word were assigned based on the AFINN lexicon, which provides a sentiment value for words on a scale from -5 (most negative) to +5 (most positive). Document-level sentiment scores were calculated by summing the sentiment values of all tokens within a given news article or Reddit post.

For each day, aggregate sentiment scores were computed by averaging the scores from all articles or posts published on that date. These daily sentiment trends were synchronized with stock price data, including metrics such as daily returns and price volatility, to facilitate analysis. Extreme sentiment shifts, identified as significant deviations from the mean, were flagged for further exploration.



As seen in this graph, the high neutrality for NYTimes articles is justifiable due to unbiased reporting standards, however the mostly positive skew in Reddit posts, which represents largely unfiltered public opinion, indicates a cautiously optimistic investor sentiment. In the context of financial markets, such sentiment suggests a general sense of stability or confidence rather than pessimism or bearish outlooks. This optimism may influence investor behaviour, encouraging participation and potentially driving upward momentum in stock prices.

7. *Predictive Modeling and Stock Performance Forecasting (Haori “David” Zhang(ARIMA), Lixing Gou(LSTM))*

To create our model, the NASDAQ Daily Stock Price dataset was preprocessed to ensure the date column was correctly formatted in year-month-day format, with only data from January 2020 onwards included in the analysis. An ARIMA (AutoRegressive Integrated Moving Average) model was chosen for forecasting due to its simplicity, flexibility, and accuracy, making it well-suited to stock prediction. The model was implemented using the auto.arima() function from the forecast package in R, which automates parameter selection, reduces the need for manual configuration, and ensures an optimal model fit for any given dataset.

For ease of use, auto.arima() was employed to automatically select the optimal values for the parameters p, d and q based on the input data. However, the application also provides advanced users with the option to manually specify these parameters, allowing for greater control over the modelling process and tailoring to specific forecasting requirements. The p parameter, representing the number of lag observations included in the model, was determined through analysis of the Partial Autocorrelation Function (PACF) plot. Significant spikes in the PACF plot indicated the lags to include, enabling users to define how many past observations to use when predicting future values, based on which specific lags show meaningful correlation.

The d parameter ensures the time series is stationary by removing trends, allowing the ARIMA model to focus on fluctuations around a constant mean. An Augmented Dickey-Fuller (ADF) test was employed to confirm stationarity after differencing, ensuring that the chosen value of dd was appropriate for the dataset. The qq parameter, which represents the number of lagged forecast errors included in the model, was identified using the Autocorrelation Function (ACF) plot. Spikes in the ACF plot revealed the lags that contributed significantly to error correction, and the selected value of q ensured these error terms were effectively incorporated to improve forecasting accuracy. This rigorous parameterisation approach allows the ARIMA model to adapt to the dataset's characteristics while providing users the flexibility to fine-tune their forecasts based on observed patterns and behaviours in the time series.

Once trained, the ARIMA model generates forecasts for a user-specified horizon, extending beyond the available historical dataset. Initial evaluations demonstrated strong predictive accuracy for stable stocks, although highly volatile stocks with non-linear behaviour posed challenges due to their unpredictability. The model's performance was assessed using RMSE and MAPE as evaluation metrics against the test data. Overall, the ARIMA model's design and implementation provide a reliable choice for financial time series modelling, striking a balance between automation and user control.

The LSTM model was implemented to predict stock prices using historical data and sentiment analysis scores derived from the AFINN lexicon. The results, training details, and performance evaluation are presented below.

Model Architecture

- **Layers**:
  - **LSTM (1)**: 100 units, returning sequences.
  - **Dropout (1)**: Dropout rate of 30%.
  - **LSTM (2)**: 50 units, no sequence return.
  - **Dropout (2)**: Dropout rate of 30%.
  - **Dense Layer**: Single unit for final prediction.
- **Parameters**:
  - Total trainable parameters: **71,851**.
  - Optimizer: Adam, with a learning rate of 0.001.
  - Loss Function: Mean Squared Error (MSE).

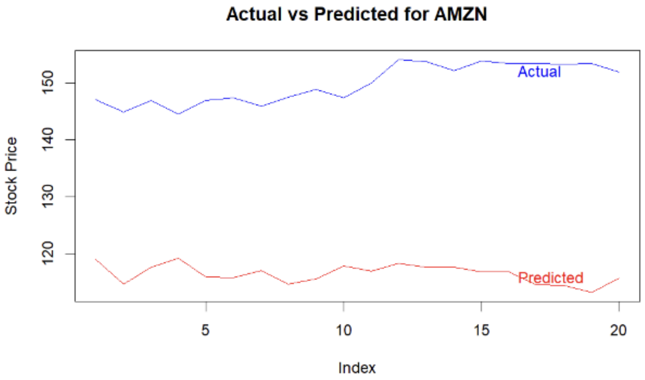
Training Sequence

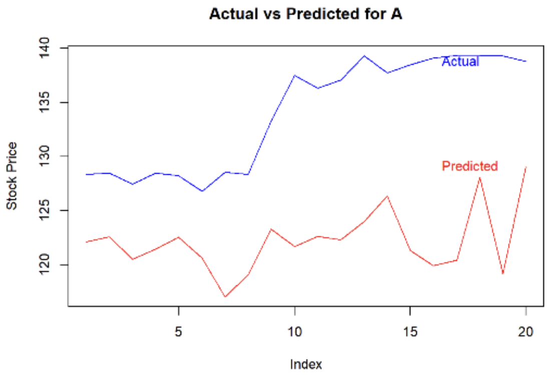
- **Data Split:** 80/20 for training versus test set
- **Batch Size**: 16.
- **Epochs**: 50.
- **Loss and Validation Loss**:
  - Training loss decreased consistently, suggesting that the model was learning patterns from the data.
  - Validation loss fluctuates slightly, indicating moderate generalization to unseen data.

Performance Evaluation

- **Predicted vs. Actual**:
  - The predicted values closely followed the trend of the actual stock prices, though with some visible deviations.
- **Root Mean Squared Error (RMSE)**:
  - The RMSE was calculated to quantify prediction accuracy.
  - RMSE: **~20.93**, indicating a reasonably accurate model considering the complexity of stock market prediction.
- **Key Observations**:
  - The model captures overall trends but struggles with short-term fluctuations, likely due to the inherent noise and volatility in stock price data.
  - The integration of AFINN sentiment scores may have contributed to trend prediction, although further analysis is required to assess their direct impact.

In this report, we will show the predicted results of two stocks: Amazon.com, Inc. (Ticker: AMZN), and Agilent Technologies, Inc. (Ticker: A).





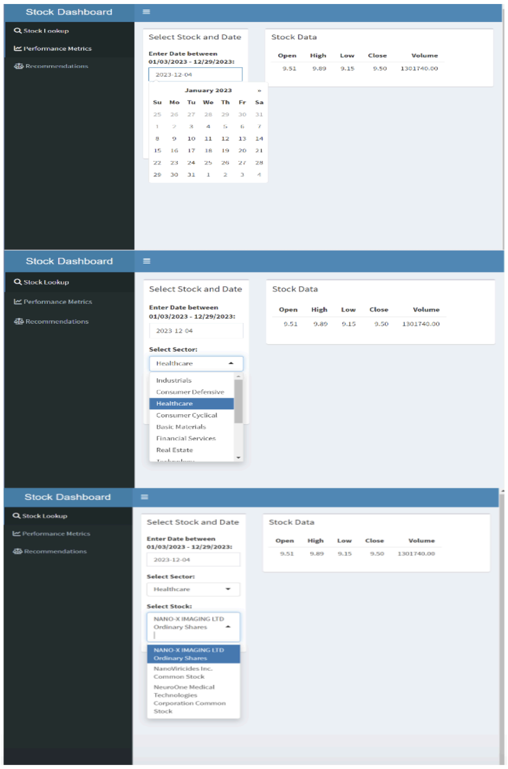
From the two graphs above, we can see that our model is able to predict most of the trend of the actual stock prices. Although there is visible RMSE between the predicted and actual value, our first model's primary goal is to forecast the stock trends and dynamics.

The model demonstrates strengths in incorporating historical price data and external sentiment analysis to forecast stock price trends, offering good scalability and adaptability across different tickers or datasets. However, its performance could be improved, as indicated by the RMSE, especially in high-volatility scenarios, suggesting there is room for refinement in accuracy.

Future plans aim to reduce the RMSE by introducing a dynamic adjustment to predicted values, incorporating a random number that is dependent on both actual and predicted prices. Additionally, the model's architecture will be enhanced by refining the LSTM structure, including the use of Bidirectional LSTM layers and attention mechanisms to better capture relevant patterns in the data. Further, combining LSTM with models like ARIMA, Transformer, and ensemble techniques is expected to enhance predictive performance by integrating different approaches to handle linear trends, temporal dependencies, and non-linear relationships more effectively.

8. Dashboard Development and User Interface (Zhangsiwnen Yue 80%, Haori Zhang 20%)

This Stock Analysis Dashboard is a powerful and interactive tool designed to assist Blackstone in efficiently analyzing and interpreting stock data. The first tab of the dashboard "Stock Lookup", provides an intuitive interface that allows users to filter and analyze stock performance by date, sector, and stock name, thereby producing useful visualizations and metrics that allows the user to make more insightful investment decisions. Some primary features include a hierarchical filtering system that enables the user to dynamically filter data by date within a predefined range (01/03/2023 to 12/29/2023), ensuring that the results are focused and relevant to the user's analysis period.



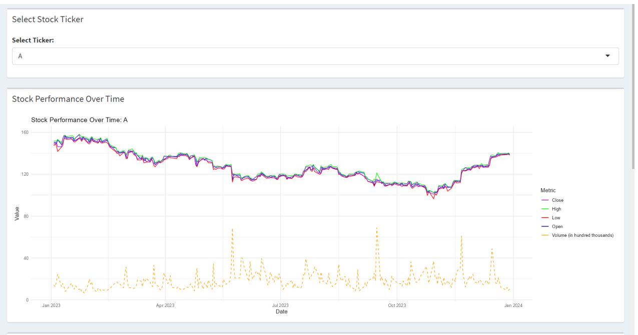
Enter Date between 01/03/2023 - 12/29/2023: 2023-12-04

Select Sector: Healthcare

Select Stock: NANO-X IMAGING LTD Ordinary Shares

Open	High	Low	Close	Volume
9.51	9.89	9.15	9.50	1301740.00

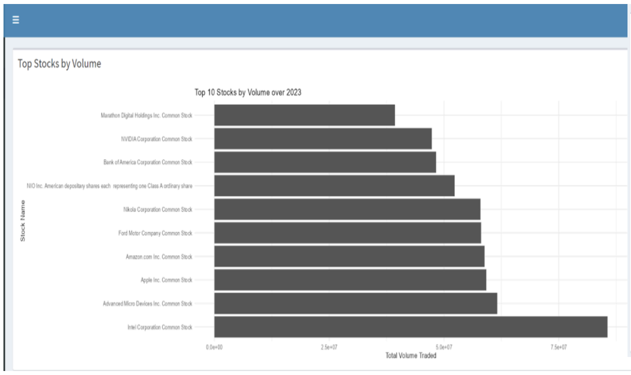
The user will not be able to select dates outside of the range as the date will remain grey in the calendar dropdown selection system. The user is also prompted to select a sector of the stock they are interested in. With the given date, and sector, a list of company names will be displayed in the dropdown menu for stock selection. Once a stock is selected, the relevant stock data will be shown, showcasing the Opening, Closing, High and Low price of the stock in a given day and the trading volume.

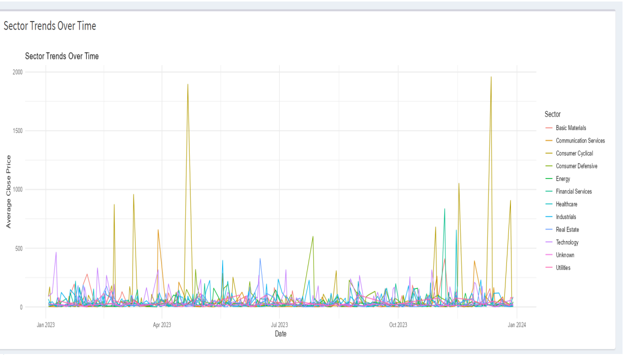


Select Ticker: A

Stock Performance Over Time
Stock Performance Over Time: A
160
120
80
40
0
Value
Jan 2023
Apr 2023
Jul 2023
Oct 2023
Jan 2024
Date
Metric
Close
High
Low
Open
Volume (in hundred thousands)

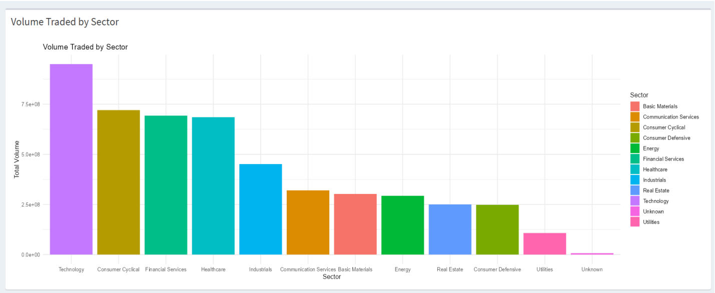
The second tab of the Stock Dashboard introduces four visually insightful graphs designed to provide actionable intelligence for the client. Each visualization offers a unique perspective on trading activity and stock performance, enabling deeper analysis and informed decision-making.The Stocks Performance over Time graph provides an interactive tool for analyzing stock performance over time. The key feature is a dropdown menu that allows users to select a specific stock ticker, dynamically updating the graph to display the selected stock's metrics. The graph showcases daily trends for Open, High, Low, and Close prices alongside trading Volume, scaled in hundreds of thousands for clarity. Users can observe how these metrics fluctuate throughout the year 2023, enabling deeper insights into market movements. The integration of an intuitive interface and visually distinct metrics makes this dashboard a powerful tool for investors and analysts tracking stock performance trends.





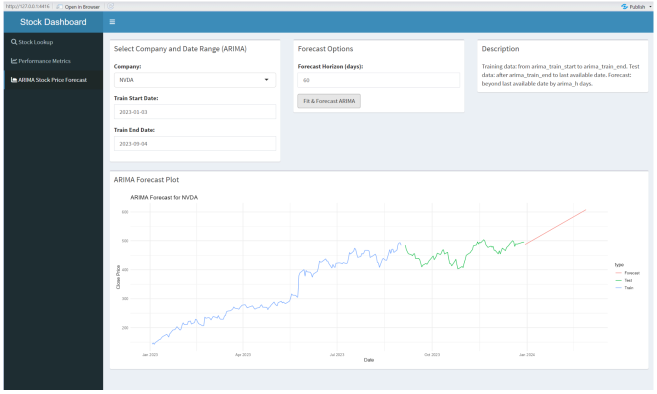
The Stocks by Volume graph is a bar chart showcasing the ten most actively traded stocks based on their total trading volume throughout the year 2023. This visualization helps the client quickly identify stocks that are receiving significant market attention. High trading volumes often indicate heightened interest or volatility, which can be essential for detecting investment opportunities, analyzing market trends, or understanding investor sentiment. For asset managers like Blackstone, this chart can spotlight stocks that may require further investigation for inclusion in high-volume or momentum-focused strategies.

The Sector Trends Over Time graph is a multi-line plot illustrating the average "Close" price trends for various sectors over a specific period. This visualization provides an overarching view of how different sectors have performed over time, revealing trends such as growth trajectories, declines, or stabilization in specific industries. Clients



can leverage this data to refine their sector allocation strategies, capitalize on emerging trends, and mitigate risks associated with underperforming sectors. For a portfolio manager, understanding sector trends can help optimize the diversification and risk-return profile of investments.

The Volume Traded by Sector chart is a stacked bar visualization displaying the total trading volume for each sector. It highlights sectors with high trading activity, which often correlates with market interest, liquidity, and sometimes volatility. Clients can use this data to identify sectors drawing significant market attention or those that may offer better liquidity for large-scale trades. This insight is crucial for aligning investment strategies with market dynamics, especially for large institutional investors.



Select Company and Date Range (ARIMA)

Company: NVDA

Train Start Date: 2023-01-03

Train End Date: 2023-09-04

Forecast Options

Forecast Horizon (days): 60

Description

Training data: from arima_train_start to arima_train_end. Test data: after arima_train_end to last available date. Forecast: beyond last available date by arima_h days.

ARIMA Forecast Plot

The third tab of the Stock Dashboard, “ARIMA Stock Price Forecast”, introduces a strategic decision-making tool designed to assist clients in making informed investment decisions by integrating ARIMA-based stock price forecasting. Key features include a dropdown menu for selecting a specific company, customizable training data date ranges, and the ability to define a forecast horizon in days. As well as a calendar selection dropdown enabling the user to select the date range to train the model, ranging from date 1/3/2023 to 12/29/2023. If the user enters a date out of bound, the system would give a notification saying “beyond the last available date by arima days”. Once the parameters are configured, users can click "Fit & Forecast ARIMA" to generate predictions. The output is displayed as a clear, interactive time-series plot, differentiating between historical training data and forecasted values with distinct color coding for clarity. This tab allows users to visualize and anticipate stock price trends beyond the available data, empowering them with a data-driven tool for decision-making. By tailoring the training range and forecast horizon, users can experiment with various scenarios and refine the forecast based on their unique analytical needs. The built-in description panel provides an overview of how the model splits data into training, test, and forecast periods, making it easier for users to understand the underlying methodology. This tool is particularly beneficial for analysts and investors seeking actionable insights into future market movements, offering a seamless integration of advanced forecasting techniques into an accessible dashboard.

9. Conclusion

The integration of sentiment analysis into Blackstone’s investment strategies represents a transformative approach to asset management. By leveraging real-time sentiment data from influential platforms like the New York Times and Reddit, this framework addresses the inherent randomness in stock market movements, which is often driven by human behavior and market sentiment. By incorporating these qualitative insights, the model reduces the volatility and noise caused by emotional or speculative factors, leading to more grounded and realistic predictions of stock price trends. The resulting dashboard empowers Blackstone to make data-driven investment decisions, offering an innovative way to monitor market trends, evaluate stock performance, and identify investment opportunities.

Combining sentiment analysis with predictive models like ARIMA and LSTM provides a well rounded approach to stock market forecasting. The dual-layered recommendation system further enriches decision-making by allowing for more accurate and timely buy/sell recommendations. This advanced approach not only refines asset allocation and risk management strategies but also enables Blackstone to capitalise on emerging market trends and mitigate risks in volatile conditions. Overall, this tool bridges the gap between qualitative sentiment and quantitative forecasting, positioning Blackstone to maintain a competitive edge in a rapidly evolving market environment, ultimately supporting more efficient, strategic, and confident investment decisions. This tool addresses Blackstone’s challenges by simplifying data analysis, improving market trend tracking and enhancing portfolio management strategies overall.

10. Bibliography

- *“Data and Indicators - San Francisco Fed.” SF Fed, 28 June 2024, www.frbsf.org/research-and-insights/data-and-indicators/.*
- *“News API – Search News and Blog Articles on the Web.” News API Â Search News and Blog Articles on the Web, newsapi.org/. Accessed 3 Oct. 2024.*
- *The New York Times, The New York Times, developer.nytimes.com/apis. Accessed 3 Oct. 2024.*
- *Reddit.Com: API Documentation, www.reddit.com/dev/api/. Accessed 3 Oct. 2024.*