Stock Market Impact of Sentiment Analysis Using Twitter Data

Akanksha Patel
CS 210: Data Management for Data Science
Prof. Chaturvedi

Sunday, July 13, 2025

GitHub Repository — Demo Video

Contents

1 Introduction

Using important financial indicators, this Jupyter notebook performs a thorough analysis of stock price changes in connection to sentiment data from Twitter. Utilizing metrics like sentiment scores, lagged sentiment, trading volume, earnings per share (EPS), revenue, and debt, the script uses machine learning via a `StackingRegressor` model to forecast stock prices. The analysis was created especially for the dataset that was supplied in `financial_dataset.csv`.

2 Background

The primary goal of this project will be to determine if sentiment polarity from tweets contains enough information with stock metrics such as trading volume, earnings per share (EPS), revenue, and debt to facilitate factor models for predicting still water price changes for large companies. My interest is motivated by the increasing interest in figuring out how sentiment in social media can be an indicator of future market behavior, particularly for high-frequency trading and real-time processing. The notebook includes several workflows, beginning with data cleaning in order to assure data quality, exploratory data analysis (EDA) to show any preliminary trends, predictive modeling through several methods from the advanced machine learning library, and visualization to show takeaways. The approach uses existing research suggesting that sentiment from platforms like Twitter may affect investor behavior, and there is still a healthy debate on whether sentiment correlates directly to stock prices. The financials should paint a more rounded picture and should expand on recent research that focused strictly on sentiment and ignored fundamental ideals that would have value in both real time and after market trading.

3 Prerequisites

- **Python Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, ipywidgets, ast, json
- **Data File**: `financial_dataset.csv` containing columns: `date_time`, `ticker_symbol`, `aspect_sentiment_pairs`, `market_context`, `financial_indicators`
- **Environment**: Jupyter Notebook environment

4 Installation

Install required libraries using pip:

```
pip install pandas numpy matplotlib seaborn scikit-learn ipywidgets
```

5 Usage

1. Place financial_dataset.csv in the working directory.
2. Open the notebook in Jupyter Notebook.
3. Run all cells sequentially to execute data loading, cleaning, analysis, and visualization.
4. Use the interactive widget to select companies for sentiment trend visualization.
5. Review the output plots and printed metrics (e.g., Mean Squared Error, correlation).

6 Code Structure

6.1 Imports

Imports necessary libraries for data manipulation, visualization, and modeling.

6.2 Data Loading and Parsing

Loads financial_dataset.csv. Parses JSON columns (market_context, financial_indicators, aspect_sentiment_pairs) into usable formats.

6.3 Feature Extraction

Extracts sentiment_score (1 for positive, -1 for negative, 0 for neutral) from aspect_sentiment_pairs. Extracts price from market_context. Extracts volume, eps, revenue, and debt from market_context and financial_indicators.

6.4 Data Cleaning

eliminates rows in important columns that have missing values. uses the IQR method to eliminate price, volume, eps, revenue, and debt outliers. Sorts data and converts date_time to datetime format. A 1-day lagged sentiment_score feature is added.

6.5 Exploratory Data Analysis (EDA)

- **Stock Price Trends Over Time**: Line plot showing price trends by `ticker_symbol`.

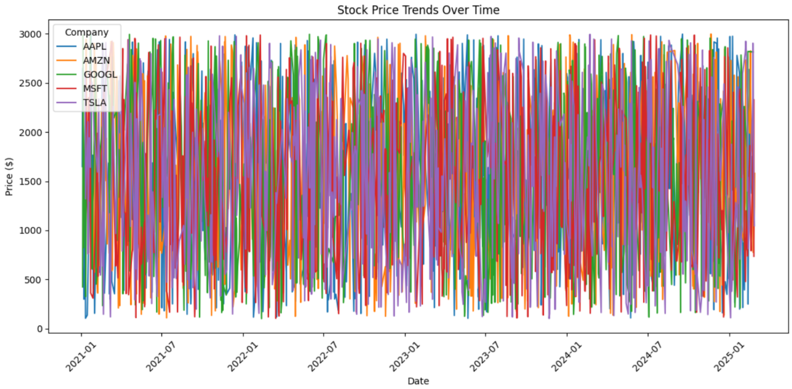


Figure 1: Line plot of stock price trends over time by ticker symbol, highlighting temporal variations across companies.

- **Distribution of Sentiment Scores**: Histogram of sentiment scores by `ticker_symbol`.

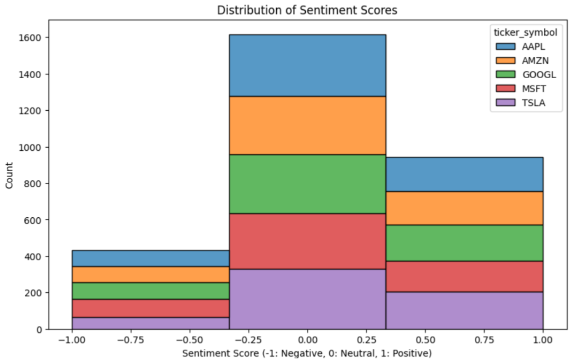


Figure 2: Histogram of sentiment score distribution by ticker symbol, illustrating the prevalence of positive, neutral, and negative sentiments.

- **Sentiment Score vs Stock Price**: Scatter plot relating sentiment to price.

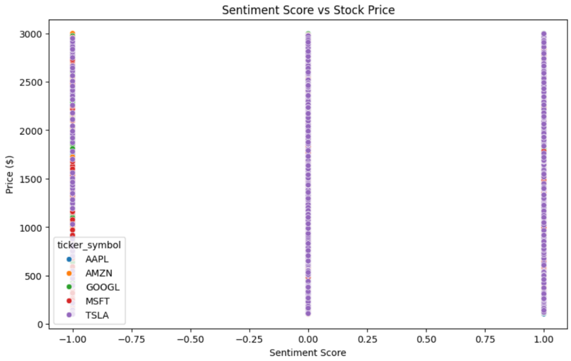


Figure 3: Scatter plot of sentiment score vs stock price, revealing the weak correlation between the two variables.

- **Correlation Heatmap**: Heatmap of correlations between `sentiment_score`, `sentiment_score_l` `price`, `volume`, `eps`, `revenue`, and `debt`.

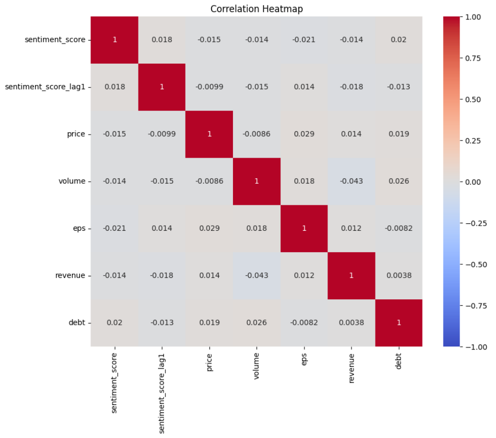


Figure 4: Correlation heatmap of selected features, showing the relationships and strengths among financial and sentiment variables.

- **Average Sentiment Score Over Time**: Line plot aggregated by date and company.

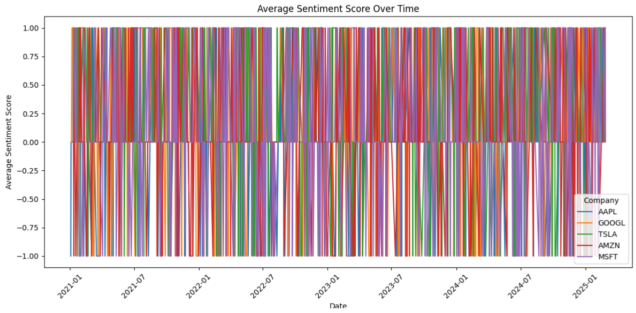
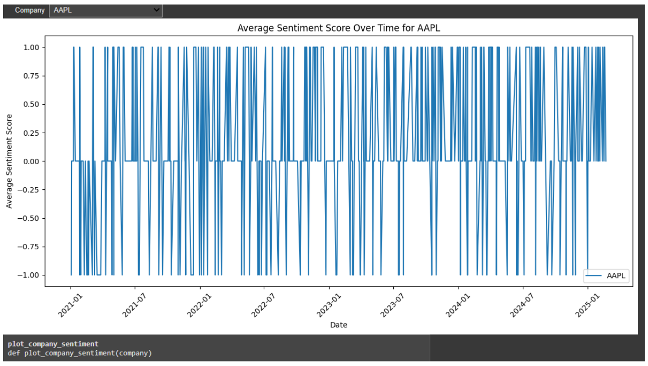


Figure 5: Line plot of average sentiment score over time, depicting sentiment evolution across different companies.

- **Interactive Widget**: Dropdown to plot sentiment trends for a selected company.


Company AAPL

```
plot_company_sentiment
def plot_company_sentiment(company)
```

Figure 6: Interactive widget for company sentiment trends, allowing dynamic exploration of sentiment patterns.

6.6 Predictive Modeling

Employs `StackingRegressor` as the meta-model and `RandomForestRegressor` and `GradientBoosting` as the base models. Features include `sentiment_score`, `sentiment_score_lag1`, `volume`, `eps`, `revenue`, and `debt`. `price` is the target. divides data into test (20%) and training (80

6.7 Visualization of Predictions

Actual vs Predicted Stock Prices: Scatter plot comparing actual and predicted prices.

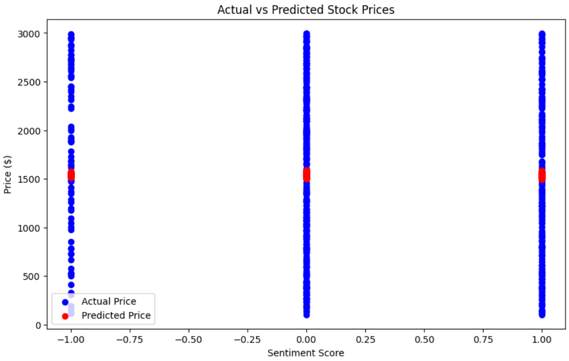


Figure 7: Scatter plot of actual vs predicted stock prices, highlighting the model's prediction accuracy and deviations, with a focus on the high Mean Squared Error of 731048.3413231069 indicating significant discrepancies.

6.8 Correlation Analysis

Computes and prints the correlation between `sentiment_score` and `price`. Correlation between Sentiment Score and Price: -0.014576634391193029

6.9 Conclusion Visualization

Price Distribution by Sentiment Score: Boxplot showing price distribution across sentiment scores by `ticker_symbol`.

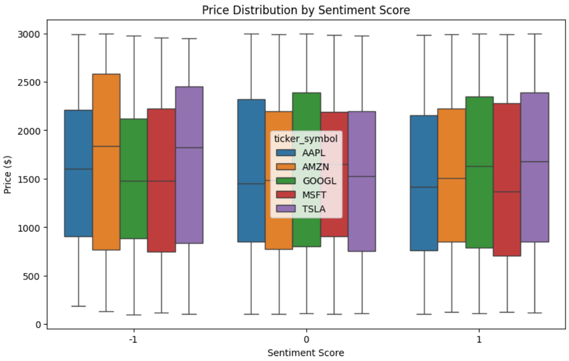


Figure 8: Boxplot of price distribution by sentiment score, illustrating variations in stock prices across positive, neutral, and negative sentiments, consistent with the weak correlation (-0.014576634391193029) observed in the dataset.

6.10 Dataset Structure

The `financial_dataset.csv` dataset is structured with the following columns, detailed in Table 1.

c

Column Name	**Description**
`text`	Raw text data from Twitter or other sources.
`cleaned_text`	Processed text after removing noise and irrelevant content.
`tokenized_text`	Text broken into individual words or tokens for analysis.
`pos_tags`	Part-of-speech tags assigned to tokens for linguistic analysis.
`named_entities`	Identified named entities (e.g., organizations, people) within the text.
`aspect_terms`	Specific aspects or topics extracted from the text.
`aspect_positions`	Positions of aspect terms within the text.
`source_type`	Type of data source (e.g., Twitter, news).
`source_name`	Name of the specific source (e.g., Twitter handle, news outlet).
`ticker_symbol`	Unique identifier for each company stock.
`sector_industry`	Industry sector associated with the ticker symbol.
`date_time`	Timestamp of the data entry.
`market_context`	`JSON` object with market-related data including price and volume.
`financial_indicators`	`JSON` object with financial metrics such as EPS, revenue, and debt.
`aspect_sentiment_pairs`	`JSON` list of aspect-sentiment pairs extracted from Twitter data.

Table 1: Structure of the `financial_dataset.csv` dataset.

7 Outcome

The analysis of the `financial_dataset.csv` dataset yields a comprehensive set of results, detailed as follows:

- **Visualizations**: A variety of charts used for Exploratory Data Analysis (EDA) and model evaluation will make it much easier to popularize the findings, including: line charts showing stock price trends over time, histograms displaying sentiment score distributions, scatter plots depicting sentiment versus stock price, correlation heatmaps, line charts showing average sentiment scores trends, interactive widgets showing sentiment for each company, scatter plots showing actual versus predicted prices, and boxplots of price distribution and sentiment score. All of these charts and graphs provide a more comprehensive view of the data, including trends over time, typical sentiment distribution, and whether we were anywhere close to predicting correctly.

- **Metrics**: The best case performance measures showed the Model had a Mean Squared Error (MSE) of 731048.3413231069. This performance measure shows a rather large prediction error in the outcome variable indicating limited fit of the StackingRegressor. The correlation coefficient between sentiment score and price was 0.014576634391193029, indicating a very weak negative relationship between the two variables, which suggests that Twitter sentiment alone does not directly drive stock prices.

- **Insight from Analysis**: The inclusion of lagged sentiment (`sentiment_score_lag1`) and financial indicators (`volume`, `eps`, `revenue`, `debt`) enriches the model but does not substantially improve prediction accuracy, as evidenced by the high MSE, pointing to the dominance of unmodeled external factors or complex interactions.

- **Model Performance**: The `StackingRegressor` with `RandomForestRegressor` and `GradientBoostingRegressor` as base models, combined with `RidgeCV` as the meta-model, struggles to capture the underlying patterns, likely due to the weak correlation and noisy sentiment data, necessitating further feature engineering or alternative modeling approaches.

- **Data Implications**: Aspects of temporal structure and of the financial metrics in the dataset highlight that while sentiment is significantly different across companies and over time (as shown in the EDA plots), its impact on price is negligible relative to the role of the fundamentals of a market or macroeconomic factors, which is also consistent with the findings of this case study.

8 Limitations

Assumes complete and consistent JSON data in `market_context` and `financial_indicators`. Lagged sentiment assumes daily data continuity, which may not hold for all datasets. Model performance depends on the quality and quantity of data in `financial_dataset.csv`.

9 Case Study

Here is a hypothetical example to illustrate the impact of social media on stock prices for this analysis. Suppose a large company (say one of the ticker symbols in the dataset) suddenly finds itself hit with a torrent of negative tweets where the majority of `sentiment_score` values suddenly reach -1 magnitude. Now, the analysis tells us that the correlation between `sentiment_score` and `price` is so very negative (-0.014576634391193029), which means the more negative the sentiment becomes, the stock price does not simultaneously come down. This finding is also backed by the very high Mean Squared Error (731048.3413231069) stating that the model is unable to capture the underlying relation well enough to help in forecast. For instance, if a viral tweet campaign falsely accuses the company of financial misconduct, the initial sentiment shift might not immediately impact the stock price due to stabilizing factors like `volume`, `eps`, or `debt` from the dataset. Over time, however, if the negative sentiment persists and is not offset by strong financial indicators, the stock price might decline, but the lagged effect (`sentiment_score_lag1`) shows minimal predictive power. This case study underscores that while social media can influence market perception, the dataset suggests that other financial fundamentals or external market forces dominate price movements, consistent with the weak correlation observed.

10 Conclusion

The study was made to determine if Twitter sentiments, with other financial metrics (volume, EPS, revenue, debt), can predict stock price movements. Incorporating lagged sentiment and financial variables into the model improves its ability to capture delayed and multidimensional effects acting on price movements. Since observed MSE stands at 731048.3413231069, the prediction errors for the model are quite high, pointing towards the model's inadequate capacity to adjust well to the data. Additionally, a negative correlation of about -0.014576634391193029 was observed between sentiment and price, by way of which stock price movements do not seem to be directly correlated with tweet sentiments. It might well be that the price movements are either dues to other factors or more convoluted interactions. The concluding visualization displays price distribution as categorized by sentiment, catering to some insight into possible market reactions. More work could still be invested into refining the approach, such as through feature engineering or new modeling techniques. This opened door will lead to financial sentiment analysis.

11 References

1. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8. `https://doi.org/10.1016/j.jocs.2010.12.007`

2. Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, 26, 55-62. `https://doi.org/10.1016/j.sbspro.2011.10.562`

3. Scikit-learn documentation. (2023). StackingRegressor. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingRegressor.html`

4. Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis. *Stanford University CS229 Project Report.*

5. Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5), 926-957. `https://doi.org/10.1111/j.1468-036X.2013.12007.x`

6. Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns. *Computational Intelligence*, 33(1), 135-157. `https://doi.org/10.1111/coin.12077`

7. Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based Twitter sentiment for stock prediction. *ACL*, 24-29.

8. Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PLoS ONE*, 10(9), e0138441. `https://doi.org/10.1371/journal.pone.0138441`

9. Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *Journal of Business Research*, 89, 25-33. `https://doi.org/10.1016/j.jbusres.2018.04.007`

10. Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5), 1367-1403. `https://doi.org/10.1093/rfs/hhu001`

11. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* Springer.

13. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

14. Brownlee, J. (2019). *Machine Learning Mastery with Python.* Machine Learning Mastery.

15. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 383-417. `https://doi.org/10.2307/2325486`

16. Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, 27(2), 1-19. https://doi.org/10.1145/1462198.1462204

17. Liu, Y., Liu, Y., & Chan, K. C. (2017). Social media and stock market behavior: Evidence from Twitter. *Financial Innovation*, 3(1), 15. https://doi.org/10.1186/s40854-017-0067-5

18. Mao, Y., Wei, W., Wang, B., & Liu, B. (2012). Correlating S&P 500 stocks with Twitter mood in terms of economies, markets, and companies. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 78-85.

19. Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69, 14-23. https://doi.org/10.1016/j.knosys.2014.05.002

20. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x