Sentiment Analysis of News Title and Its Predictive Power on Tesla's Stock Movements and Price

Group 12: Leo Liu, Mika Tanchavanich, Shanshan Wang, Wanyi Liu

Research Problem

Stock prices are determined by supply and demand; it is determined by the highest price a buyer is willing to pay and the lowest price the seller is willing to sell. For decades, researchers have been trying to determine the future value of stocks by trying to understand the factors that contribute to the balance of supply and demand. One of the crucial factors is understanding the market dynamics and sentiment. This is particularly important for Tesla, Inc.

Tesla, Inc. is a leading electric vehicle company that has been publicly traded on NASDAQ under the ticker symbol TSLA since 2010. It is one of the most prevalent and volatile companies in the market with a current YOY return of 129% and a five-year return of 1073% (Kabir, 2023). One of the reasons why TSLA value is so volatile is because it is heavily reliant on market perception and sentiment; positive and negative news, social media discussions, and public opinions are all impactful on stock prices. As such, this study aims to understand how news sentiment reflected in headlines and media coverage correlates with TSLA stock performance to understand the market dynamics and sentiment towards the stock performance.

Related Literature

The relationship between stock prices and market sentiment is a subject of considerable academic interest and practical relevance in finance. Traditionally, shareholders sought insights into their investment decisions through word of mouth and traditional media outlets, inspiring the first study on the automatic prediction of stock prices using machine learning regression models on newspaper text that date back to 1993 (Refenes). Recent research, such as that conducted by Smith in 2022, has shed light on the impact of tweets on social media platform X.

In 2023, Palomo delves into using tweets to train a random forest classifier model, achieving a 91% accuracy rate in predicting stock price. Using weighted classification, which accounts for the effect between different authors, her results yielded an 82% accuracy (Palomo, 2023).

Many other researchers have conducted their prediction methods including Bujari, Furini, and Laina. In their 2017 study, they created a custom model extracting features and analyzing the sentiment of tweets (Bujari, Furini, Laina). Nevertheless, it's important to note that a one-size-fits-all model doesn't exist, as prediction accuracy fluctuates significantly across different stocks over a 70-day period.

While tweets seem to give insight into stock prices, our study aims to use a more comprehensive view of TSLA by looking at news headlines. The headlines in the dataset include insights from investment and consulting firms, financial reports, and as well as social media. Moreover, our research explores various hypothetical targeting words, including those associated with price increases analyst ratings, earning reports, capital structure changes, and mergers and acquisitions. (Kabir, 2023) Conversely, factors like political and economic uncertainties and bad earning reports may cause a price decrease (Beers, 2021).

Research Questions to be Examined

This study aims to address several key research questions that revolve around Telsa-related news headlines and the corresponding impact on TSLA stock prices over the past year. Firstly, we plan to delve into the potential causal relationship between specific news events, sentiment expressed in the headlines, and movement in stock prices by exploring the questions:

1. Can sentiment analysis of news headlines be used as a predictive indicator for short-term stock price movements?

2. How does sentiment in Tesla-related news headlines correlate with stock price and stock price movement?

Second, we plan to investigate the best method to predict stock prices based on news titles by running different types of models.This led to our research question:

3. What is the best way to predict short-term stock price?

Data

To explore the research questions, we will use two datasets. The first data set is the historical stock prices of TSLA (table name *TSLA*). This data was acquired from Yahoo Finance. This dataset spans a one-year timeframe and includes the date, the opening price, high price, low price, closing price, adjusted closing price, and trading volume, and is recorded daily.

The second data set is comprised of news headlines related to TSLA and the published dates of each article (table name *news_title)*. This list is on a webpage compiled by Nasdaq, a financial services corporation that owns and operates multiple stock exchanges globally.

By comparing these two datasets, we aim to uncover patterns and correlations between the sentiment conveyed in news articles and the fluctuations in TSLA prices over the past year.

Data Preparation

To prepare the first dataset, we downloaded the daily historical prices of TSLA for the previous year (28 February 2023 - 27 February 2024). To prepare our second dataset, we initially searched for a dataset that contains news titles associated with Tesla, Inc. from past years. However, we found limited resources that are not available for our project. Therefore, we decided to scrape news titles from a financial website. Using the libraries `Requests' and `BeatifulSoup,` we wrote a Python script to crawl from nasdaq.com. This approach did not work because, upon further investigation, we discovered that the website is dynamic and that the news headline data is encapsulated in JavaScript, which prevents direct access from the HTML source code. As such, we leverage the `selenium` package to facilitate our access to JavaScript-generated data. The website only displays 5 news headlines at a time and has dynamic pagination integrated for accessing data on the next page. We sent a click signal using methods from `selenium` to update the data we needed to scrape. Eventually, we merged all the data we collected from the webpage into a data frame and exported the data frame as a .csv file.

These two datasets were well formatted with no missing values and little inconsistencies. To clean the dataset, we standardized the date formats. In the *tesla_news* table, some dates we recorded as ‘x hours ago’ or ‘x days ago’. Using the date_string function, we standardized the date format and chronologically organized the news headlines.

Next, we created a new dataframe, *df*, which combines the date, news title, and closing stock price. For initial data exploration, we aimed to explore the most common words in the news title. By tokenizing the news title to analyze each title by individual words, we were able to discover context-specific stopwords that were prevalent in the news title. For example, we removed words like “tesla”, “tsla”, “stock”, “stocks”, “2023”, and “2024” and tidytext::stopwords in order to focus the sentiment analysis on more emotive language that might better capture the mood or opinion of the view.

word <chr>	count <int>
us	887
3	535
ev	514
buy	463
earnings	383
stockswall	301
data	272
st	262
investors	199
etf	192

Figure 1. Top 10 most common words after removing stop words

Next, we preformed vectorization. This step was crucial in cleaning and standardizing the text. We converted to lowercase, removed punctuation, stripe whitespace, and removed stopwords. This step will help us enhance the accuracy of sentiment analysis. This technique enables similar tokens such as “Earning” and “earning” to be one word. Doing this allows us to reduce redundancy. Also, removing stop words and our own stop words eliminates the effect of noise.

Our next step was to preform sentiment analysis. We used the NRC and Loughran-McDonald lexicon to categorize words into different sentiments, which we would later use to quantify the emotional content of the news title to the stock prices. The Loughran-McDonald lexicon is tailored for financial terms and sentiment analysis in the domain of finance. Overall, the top emotions of all the news titles were positive, negative, and anticipation/uncertainty.

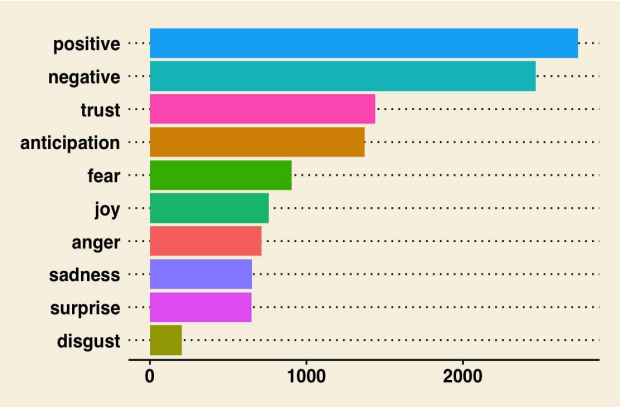

Figure 2. Top sentiment using NRC Emotion Lexicon

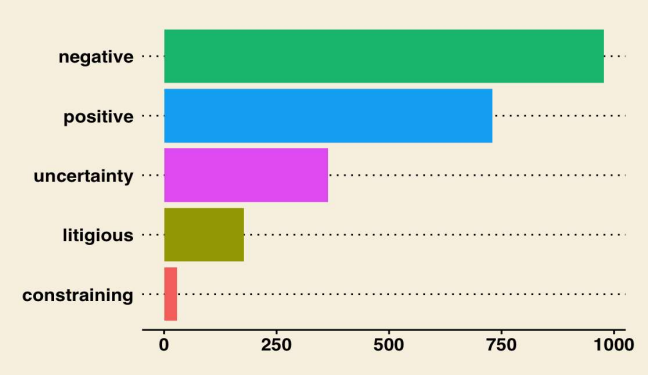

Figure 3. Top sentiment using Loughran-McDonald Lexicon

Then, we grouped the sentiment by date and joined it with the stock price under the dataframe

Date	Open	High	Low	Close	Adj.Close	Volume	anger	anticipation	disgust	fear	joy	negative
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2023-02-28	210.59	211.23	203.75	205.71	205.71	153144900	1	2	1	1	1	4
2023-03-01	206.21	207.20	198.52	202.77	202.77	156852800	2	6	0	1	2	10
2023-03-02	186.74	193.75	186.01	190.90	190.90	181500700	5	14	1	6	7	15
2023-03-03	194.80	200.48	192.88	197.79	197.79	154193300	2	4	0	3	2	5
2023-03-06	198.54	198.60	192.30	193.81	193.81	128100100	0	2	1	2	3	4
2023-03-07	191.38	194.20	186.10	187.71	187.71	148125800	2	5	0	3	2	9
2023-03-08	185.04	186.50	180.00	182.00	182.00	151897800	1	5	0	2	3	4
2023-03-09	180.25	185.18	172.51	172.92	172.92	170023800	3	3	1	2	2	9
2023-03-10	175.13	178.29	168.44	173.44	173.44	191488900	2	2	1	3	1	5
2023-03-13	167.46	177.35	163.91	174.48	174.48	167790300	3	2	0	5	1	7
2023-03-14	177.31	183.80	177.14	183.26	183.26	143717900	7	2	0	6	3	18
2023-03-15	180.80	182.34	176.03	180.45	180.45	145995600	1	2	0	0	1	7
2023-03-16	180.37	185.81	178.84	184.13	184.13	121136800	1	1	0	0	1	2
2023-03-17	184.52	186.22	177.33	180.13	180.13	133197100	2	0	0	0	1	4
2023-03-20	178.08	186.44	176.35	183.25	183.25	129684400	2	4	0	0	2	7
2023-03-21	188.28	198.00	188.04	197.58	197.58	153391400	2	3	2	2	4	5
2023-03-22	199.30	200.66	190.95	191.15	191.15	150376400	2	2	0	0	0	7
2023-03-23	195.26	199.31	188.65	192.22	192.22	144193900	1	2	0	1	1	2
2023-03-24	191.65	192.36	187.15	190.41	190.41	116312400	1	4	0	0	3	5
2023-03-27	194.42	197.39	189.94	191.81	191.81	120851600	3	5	0	3	4	8
2023-03-28	192.00	192.35	185.43	189.19	189.19	98654600	0	2	0	0	1	1
2023-03-29	193.13	195.29	189.44	193.88	193.88	123660000	4	5	1	3	3	12

Figure 4. Preview of *tsla_sentiment_nrc*.

We aimed the explore the correlation between specific sentiments and stock prices. However, the correlation observed was notably low, indicating several key insights to keep in mind for the next steps: Firstly, other variables may be more impactful on stock prices. Secondly, non-linear models may yield better result. Third, more intricate models may yield better results. We kept these findings in mind while choosing our models.

sentiment	Correlation with r...	p
<chr>	<dbl>	<chr>
anger	0.11	not significant
anticipation	0.13	p < 0.05
disgust	0.01	not significant
fear	0.17	p < 0.05
joy	0.09	not significant
negative	0.16	p < 0.05
positive	0.06	not significant
sadness	0.00	not significant
surprise	0.08	not significant
trust	-0.06	not significant

Figure 5. Correlation between sentiment (NRC) and stock price

Exploration the relationship between stock price movement and sentiment

In order to answer our first and second research question, we used feature engineering to create a new column *difference* indicating the price difference between the current closing price and previous day closing price. Then, we joined this to the sentiment and looked at the correlation and conducted a logistic regression. Our confusion matrix, as seen in figure 6, yielded an accuracy of 0.6 and a p-value of 0.2816. This model tends to be skewed towards predicting price increases. This may have to do with the fact that the time period had a bull market with 28% YOY increase in the S&P 500 and 42% YOY increase in NASDAQ during the period 28 February 2023 - 27 February 2024.

	Reference	
Prediction	Increase	Decrease
Increase	7	4
Decrease	26	38

Figure 6. Confusion matrix of predicted and reference price movement through logistic

Predicting short term stock price

Following this, we split the data into a train and test dataset, with 70% of the total number of rows being selected for the training datatset. We conducted a total of 10 different models in order to find the best method to assess the strength and nature of the relationship between sentiment and stock price movement. To measure the performance of the model, we will be using the root mean square error (RMSE) shown in table 1. The RMSE measures the differences between the predicted values and actual values.

Model	RMSE	Additional notes
NRC Sentiment Analysis		
Linear Regression Model	35.13927	Poor performance, as predicted. Led us to try Random Forest and XGBoost.
Random Forest Model	34.4228	Suggests a slight improvement over linear regression.
XGBoost Model	**2.429646**	Shows a significant improvement; built-in importance score helps to identify most influential features.
Loughran-McDonald Sentiment Analysis (Lexicon adjusted for financial terms)		
Regression Model Testing for Sentiment Analysis	35.15415	Similar to the NRC linear regression model, indicating consistent performance across different sentiment lexicons.
Random Forest Model	35.93417	This is slightly worse than its linear regression counterpart, which could be due to overfitting or the model's sensitivity to the type of features used.
XGBoost Model	**1.932149**	XGBoost has the best RMSE, reinforcing the strength of advanced ensemble methods in predictive accuracy.
Latent Semantic Analysis for Topic Modelling (Provides insights into underlying topics or themes within the text, but it is difficult to capture non-linear patterns)		
Topic Model based on Latent Semantic Analysis (LSA)	37.20463	This model identifies underlying topics in news titles, which are then used to predict stock prices. The slightly higher RMSE suggests less predictive power compared to sentiment analysis.
Random Forest (LSA)	33.99159	Perform better than LSA
XGBoost (LSA)	**11.58**	Best RMSE for LSA, similar to other analyses.

Table 1. Summary of models used and RMSE

Across the different types of models, XGBoost had the lowest RMSE. This reinforces the strength of advance ensemble methods in predictive accuracy of this context.

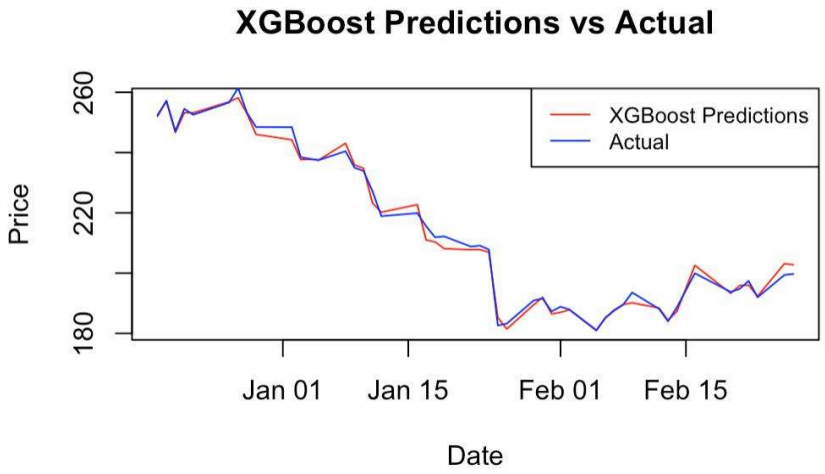

Figure 5. XGBoost (Loughran-McDonald Sentiment Analysis) predictions vs actual prices for our best model.

Conclusion

To answer the questions regarding the predictive power of sentiment analysis and the correlation between sentiment in Tesla-related news headlines and stock price movements, our study yielded insightful findings. However, we encountered limitations that warrant further investigation and refinement of our methodology. Regarding the questions

1. Can sentiment analysis of news headlines be used as a predictive indicator for short-term stock price movements?

2. How does sentiment in Tesla-related news headlines correlate with stock price and stock price movement?

We discovered that sentiment analysis itself cannot be used alone as a predictive indicator for short-term stock price movement. By exploring the relationship between sentiment and price movement (increase of decrease) using logistic regression, our model yielded 0.6 accuracy and a p-value of 0.28. This suggests that we do not have enough evidence to conclude that sentiment emotion can predict movement.

For the last question:

3. What is the best way to predict short-term stock price?

We found that the best way to predict stock price was through the XGBoost model. It produced the lowest RMSE at 1.9321. Notably, the Loughran-McDonald Sentiment Analysis showed the best prediction due to its tailored adaptation for finance-related applications.

Recognizing the limitations encountered, we outline areas for further research to enhance the validity of our findings and refine our predictive models. Firstly, we would like to expand our dataset. We plan to do this by expanding the time range beyond the one-year period as well as gathering more textual data by scraping other platforms like X where many individuals share their opinions on the company outlook. We can also explore if certain news outlet or individual's opinions hold more importance over others.

Secondly, we would like to investigate the effect of time lags. We believe that accounting for time lags may increase correlation will account for the time it takes for investors to have a improved information transparency. We plan to account for delays in information dissemination by looking at varying lag periods ranging from one to five periods.

We would also like to consider other variables that may affect stock price. One factor we believe will have a strong correlation is the performance of the market. We can compare and account for price movements in NASDAQ, S&P 500, or Dow Jones Industrial Average. These indices reflect the overall market perception and sentiment, which can have a significant impact on individual stock prices.

While our study offers insight into relationship between stock price and news headline sentiment, we acknowledge the need for further research by addressing the identified limitations and refining our predictive models.

References

1. Beers, Brian. “How the News Affects Stock Prices.” *Investopedia*, 30 9 2021, https://www.investopedia.com/ask/answers/155.asp . Accessed 29 February 2024.

2. Palomo, Chirstian. “Tweet Sentiment Analysis to Predict Stock Market.” *Stanford NLP*, 2023, https://web.stanford.edu/class/cs224n/final-reports/final-report-170049613.pdf . Accessed 29 2 2024.

3. MacDonald, Moira, et al. *Quantifying the impact of news on stock price changes*, 8 2022, https://www-2.rotman.utoronto.ca/insightshub/finance-investing-accounting/news-stock-swings. Accessed 29 February 2024.

4. Smith, Stephen, and Anthony O'Hare. “Comparing traditional news and social media with stock price movements; which comes first, the news or the price change? - Journal of Big Data.” *Journal of Big Data*, 28 April 2022, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00591-6. Accessed 29 February 2024.

5. Bujari, Armir, et al. “On using cashtags to predict companies stock trends.” *IEEE Xplore*, IEEE, 2017, https://ieeexplore.ieee.org/abstract/document/7983075 . Accessed 29 February 2024.

6. Yahoo Finance. “Tesla, Inc. (TSLA) Stock Historical Prices & Data.” *Yahoo Finance*, https://finance.yahoo.com/quote/TSLA/history . Accessed 29 February 2024.

7. NASDAQ. “TSLA News Headlines” *NASDAQ*, https://www.nasdaq.com/market-activity/stocks/tsla/news-headlines . Accessed 29 February 2024.

8. Kabir, Usman. “16 Most Volatile Stocks To Buy Now.” *Yahoo Finance*, 30 December 2023, https://finance.yahoo.com/news/16-most-volatile-stocks-buy-141732618.html. Accessed 29 February 2024.