**Sentiment Analysis in the Financial Market**

Ricardo Gehrke

Lawrence Technological University

INT 6303 – Intro to Social Media Data Analytics

Dr. Mazin Al Hamando

December 13, 2024

**Sentiment Analysis in the Financial Market** - Utilizing sentiment analysis to understand and predict behaviors of Tesla Inc. stock market price from September-30-2021, to September-28-2022

**Introduction**

Stock price fluctuations are not simple to predict. Staying up to date with news, current events, and product releases is a way to monitor stock price changes, once the Efficient-Market Hypothesis (EMH) states that these fluctuations are largely influenced by new information and will behave accordingly, making it impossible for investors to perfectly formulate predictions (Sewell, 2011). For that reason, finding a resource that can reflect moods and sentiment from a sample of the population as a reaction from news seems to be an outstanding tool to understand and build predictions. The facility of sharing information in social media transformed it into a bridge to understand public sentiment towards factors that can impact stock price changes.

Tesla Inc., an American company dedicated to creating sustainable and innovative products, is one of the powerhouses in the stock market. According to YCharts, Tesla grew 1529.3% in the last 5 years of appearing in the stock market - stocks analyzed from September 30, 2019, to September 30, 2024 (YCharts, 2024). This magnitude of growth means that the company offers a low risk and a high probability of return on investment.

**Objectives**

The general objective of this paper is to understand how sentiment analysis can contribute to stock market prediction by applying different text mining methods. More specifically, applying one rule-based algorithm and one supervised machine learning algorithm to obtain hands-on experience with different approaches. For last, this study has also as a specific goal to develop a prediction model that presents an accuracy of above 50% of success.

To keep the study simple and straightforward, the two main metrics utilized in this paper will be correlation and accuracy. The first one will be important to understand the relationship between sentiment score and stock price change, which may validate the alignment of sentiment score and stock price behavior. Next, accuracy will be used to evaluate how the prediction model performs.

**Related work**

A few studies were fundamental for this paper. Bollen, Mao, and Zeng (2010) mention that accuracy can be significantly improved when taking into account some specific emotions. Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016) broke down the steps for sentiment analysis, and utilized Random Forest and logistic regression for sentiment analysis. Jishag, A. C., Athira, A. P., Shailaja, M., & Thara, S. (2020) also used R to create prediction models for the stock market, including naive Bayes. Takawira, O., & Mwamba, J. W. M. (2021) also utilized naive Bayes for a prediction model, emphasizing its focus on treating each variable independently. Ranco G, Aleksovski D, Caldarelli G, Grčar M, Mozetič I (2015) applies lag days to understand the effect of Tweeter in stock prices variations.

**Data collection and preprocessing**

The dataset used was retrieved from Kaggle, named *Stock Tweets for Sentiment Analysis and Prediction*. This dataset has 80,794 tweets, with 37,422 mentioning Tesla. Some dates have a significantly higher number of tweets available than others, for example 2021-10-25 has 462 tweets recorded, while 2022-09-05 has 35 tweets recorded. This is a discrepancy of this project that will not be considered. In addition, even though the interval of time corresponds to almost 1 entire year, the number of unique dates available is 252. One of the reasons is because stock prices are not available during weekends. The method used to clean and select only those dates that appeared in the dataset and in the price list time frame was combining the two spreadsheets simply according to dates available. This filters out the days that do not appear in one of the two sets. Prior to that, the data manipulation step by step consisted of:

Subset and corpus: create a subset of the data frame. This would limit our work only to the rows that contain the Tesla tweets. Next, the subset is transformed into a corpus, allowing us to perform further cleaning and analysis in the text.

Cleaning text: subsequently, several cleaning tools were applied to the corpus. First, the text was converted to lowercase, ensuring words that may appear twice but without consistent case forms are treated the same. Next, stop words were removed. Those are terms that do not add sentiment to the analysis, such as *the* and *and*. In addition, characters such as punctuation, numbers, and extra whitespace were also removed.

Document Term Matrix and sparsity: after removing characters that will not aggregate value to the future sentiment analysis, the next process is to tokenize the corpus into a document term matrix (DTM). The DTM consists of a table that numbers each tweet as a row and creates a column for each term that was not discarded in the cleaning process. After creating so, it is important to evaluate the sparsity of the matrix. Sparsity refers to the number of zeros that appear in the DTM. Each zero means that the term did not appear in a tweet. By performing the following formula, *Sparsity = (1 - Total Number of Entries / Number of Zero Entries​)\*100*, the sparsity of the matrix was observed as 99.97%, indicating a high one. This value implies that the DTM is composed mainly of zeros. Shaila Miranda (2008) demonstrates that such high sparsity should be lowered to avoid the risk of rarely occurring terms to impact our analysis. Therefore, reaching a lower sparsity should be beneficial for further analysis. Thus, tools were utilized to reduce this sparsity, such as considering only terms that appear in 99% of the corpus. Figure 1 represents the code that manipulated the data:

Figure 1: lowering sparsity



Through the code above, sparsity was reduced to 97.26%, which should present improvement in future analysis. For reference, with 99.97% sparsity, the DTM presented 53,580 terms, while with 97.26% sparsity it presented 207 terms.

**Sentiment Analysis and Prediction Model**

Next step is to begin the sentiment analysis by classifying each tweet as positive, negative, or neutral. The approach used to perform the sentiment analysis was the *get\_sentiment* function from the syuzhet library. This function compares each word of each tweet to a list of dictionaries of words, classifying them. In the function’s default model, it scores each tweet, and depending on the score the tweet is classified as positive or negative. It is a rule-based method because it follows a predefined set of rules, differently from a machine learning technique, which learns from the data. After applying the algorithm, the average sentiment score was calculated through the following formula:

Average Sentiment = ∑Sentiment Scores​ / Total Number of Tweets

We were left with a table with the columns date, open, high, low, close, volume, adjusted, avg sentiment, positive count, negative count, neutral count. Being positive, negative, and neutral counts the count of how many of each tweet appear in the day with the average sentiment as well.

Another option that could have been used was natural language processing. The rule-based algorithm was used instead because we wanted to simply understand if the tweets were positive, negative, or neutral. NLP gives us an analysis corresponding to a higher granular level, classifying the tweets into additional emotions such as "anger", "fear", "joy" using get\_nrc\_sentiment() function from syuzhet.

**Correlation of price change and average sentiment**

To understand if there is correlation between stock prices and sentiment, we start by understanding what the fluctuation of each day was, named as price\_change and creating a new column with this attribute. Now we have an average sentiment for each day, and a percentage of change for each day, which is calculated from the closing price of the previous and current days. Price Change=(Previous PriceCurrent Price−Previous Price​)×100

After returning a correlation of 0.171978, the next step would be to manipulate the data so that we could increase this correlation before continuing to the machine learning prediction. Histogram shows that several tweets are close to neutral, that is, the bars between -1 and 1 not only are closely equivalent, but they also correspond to a large part of the data.

Figure 2:

A graph of a histogram

Description automatically generated

With that, fine-tuning sentiment score range could be a solution. Because most of the tweets range between -1 and 1, finding a way to value the stronger sentiment tweets could lead to a higher correlation. In more detail, if tweets that demonstrate a more extreme sentiment are given higher sentiment value, this may create a higher correlation with the stock market price change. After fine tuning, the number of neutral tweets increased. We also multiplied by 2 the value of strong positives and strong negatives, creating an average weighted sentiment score. Figure 3 shows the new sentiment label, in comparison to Figure 4, which brings the original sentiment labels:

Figure 3:



Figure 4:



The new correlation, however, went down to 0.1569316.

Next strategy that was used to attempt to increase correlation between the tweet’s sentiment score and the fluctuation of the stock prices was to apply a lag in the price change compared to the tweet sentiment scores. Ranco et al. (2015) applies 1 to 5 days of lag to understand how different moods affect the stock prices. Thus, for this project, it was decided to use 1 day of lag to observe if the correlation would be positively impacted. However, the correlation dropped to 0.03652011.

**Prediction model**

1. Because the attempts to find a higher correlation failed, the next step of applying a prediction model will be used with the steps that led the correlation to a 0.171978. The prediction model utilized was the naive Bayes model, also used in Jishag et *al*., 2020. Thara article. Naive Bayes is a supervised machine learning model that can be used to predict if the stock market price will increase or decrease, based on the sentiment of the tweets. It is a classification technique based on Bayes’ Theorem with an independence assumption among predictors, according to Analytics Vidhya (2024). In this project, however, the model will simply learn from the average sentiment score and the price change fluctuation, which was simply relabeled to increase or decrease. After applying the algorithm, the accuracy of the first attempt to use the naive Bayes was 59%.

Figure 5:

A graph of blue and pink bars

Description automatically generated

Figure 5: Additional metrics

A screenshot of a computer screen

Description automatically generated

1. Because naive Bayes considers each variable independently, correlation is not necessarily required for a high accuracy according to Oliver Takawira (2021). Due to that reason, it was important to apply the machine learning method to understand if a lower correlation would affect the accuracy. This time, the average weighted sentiment score was used to train the model. When reapplying the naive Bayes model with the average weighted sentiment scores, the accuracy decreases to 49%.
2. For last, the naive bayes method was also tested with the third correlation of approximately 0.04. This time, even though the naive Bayes model presented an accuracy of 53%, it learned only to choose the “increase” option. Because 53% of the time the price change increased, it demonstrated a higher success than the average weighted model, but in an invalid way. For that reason, this step was considered invalid. The final scheme of results is as follows:

A diagram of a graph

Description automatically generated

**Results and discussion**

It is possible to conclude this paper in two ways. Practically speaking, the results of the metrics in this study can be considered of limited success. Different techniques were applied to obtain a higher correlation, such as adjusting sentiment weights and applying lag days to cover delayed effects, but the correlation between tweet sentiment and stock price changes was never high. When utilizing the machine learning model naive Bayes, the accuracy of 59% was modest, since studies say that it is indeed difficult to surpass 75% of accuracy, for example. In general, this indicates that sentiment analysis, although important and valuable to understand the public mood, is not enough to predict stock prices with high accuracy due to the complexities and volatility of the financial market (YCharts, 2024). In addition, even though the dataset presented a high amount of tweets, for machine learning purposes it can be considered small. The one-year time frame gives us a large amount of tweets, but the number of stock price days may limit the capacity of the model to understand the behavior of the stock market, lacking in diversity and volume. Another improvement that can be considered is to add different moods in addition to only classifying tweets as positive or negative (Bollen *et al*., 2010). Next, stock prices are influenced by external factors that may not be reflected in the sentiment of social media posts. These limitations highlight the need for more extensive datasets and perhaps more complex models to improve the accuracy of predictions (Miranda, 2024).

The side of the conclusion, however, can consider this project as a success. One rule-based algorithm (get\_sentiments function from R language) and a supervised machine learning method were applied to obtain sentiment scores and develop a model that had a prediction accuracy of 59%, satisfying the general and specific goals that were set for this study.

Reference:

Bollen, J., Mao, H., & Zeng, X. (2010). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8. [https://doi.org/10.1016/j.jocs.2010.12.007](https://doi.org/10.1016/j.jocs.2010.12.007). Retrieved from: [https://www.sciencedirect.com/science/article/pii/S187775031100007X](https://www.sciencedirect.com/science/article/pii/S187775031100007X)

Equinxx [Kaggle dataset]. (2024, December 10). Stock Tweets for Sentiment Analysis and Prediction. Retrieved from [https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction](https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction)

Jishag, A. C., Athira, A. P., Shailaja, M., & Thara, S. (2020). Predicting the stock market behavior using historic data analysis and news sentiment analysis in R. In First International Conference on Sustainable Technologies for Computational Intelligence (pp. 717-728). Springer, Singapore. [https://doi.org/10.1007/978-981-15-0029-9\\_56](https://doi.org/10.1007/978-981-15-0029-9\_56). Retrieved from: [https://www.researchgate.net/publication/336988933\\_Predicting\\_the\\_Stock\\_Market\\_Behavior\\_Using\\_Historic\\_Data\\_Analysis\\_and\\_News\\_Sentiment\\_Analysis\\_in\\_R](https://www.researchgate.net/publication/336988933\_Predicting\_the\_Stock\_Market\_Behavior\_Using\_Historic\_Data\_Analysis\_and\_News\_Sentiment\_Analysis\_in\_R)

Miranda, S. (2018). Social Analytics: Network and Text Methods with NodeXL and R. O'Reilly Media, Inc.

Murphy, K. P. (2017, September). Naive Bayes Explained [Blog post]. Analytics Vidhya. Retrieved from [https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/](https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 2016 IEEE International Conference on Computational Intelligence and Communication Networks (CICN), 1-6. Retrieved from: [https://ieeexplore.ieee.org/abstract/document/7955659](https://ieeexplore.ieee.org/abstract/document/7955659)

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](https://doi.org/10.1371/journal.pone.0138441). Retrieved from [https://journals.plos.org/plosone/article/citation?id=10.1371/journal.pone.0138441](https://journals.plos.org/plosone/article/citation?id=10.1371/journal.pone.0138441)

Sewell, M. (2011, January 20). History of the Efficient Market Hypothesis (Research Note RN/11/04). Retrieved from [http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/images/Research\\_Student\\_Information/RN\\_11\\_04.pdf](http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/images/Research\_Student\_Information/RN\_11\_04.pdf)

Takawira, O., & Mwamba, J. W. M. (2021). Determinants of Sovereign Credit Ratings: An Application of the Naïve Bayes Classifier. Eurasian Journal of Economics and Finance, 8(4), 279-299. [https://doi.org/10.15604/ejef.2020.08.04.008](https://doi.org/10.15604/ejef.2020.08.04.008)

YCharts. (2024, October 4). The 10 best performing stocks in the last 25 years. YCharts Stock Screener. Retrieved from [https://get.ycharts.com/resources/blog/the-best-performing-stocks/](https://get.ycharts.com/resources/blog/the-best-performing-stocks/)