Shuju (Jeffery) Sun
DSCI 510

DSCI 510: Final Project Report

1 Introduction

The digital transformation has revolutionized dissemination of financial news, where information is shared almost instantly—a significant shift from the past when the spread of news was through traditional channels in a much slower pace. The project idea is inspired by that many financial reports are now automatically produced from datasets and earnings calls. Employing data science methods such as Natural Language Processing (NLP), data extraction, and visualization, the project aims to derive insights for investment, serving as a tool for financial firms and individual traders to inform their market strategies, including sentiment analysis from news headlines and a brief visualization for decision-making.

The research purpose of the project is to discover the relationship between stock price data and the sentiment score derived from news headlines related to the stock over a period. I have collected Tesla (ticker: TSLA) stock price data and 400 news data related to TSLA from Seeking Alpha through API as well. Once I finish cleaning up the datasets, I have briefly described the data by visualizing them and conducted NLP on the news titles. Lastly, I have processed a statistical analysis called Granger Causality Test on both the stock price and sentiment score over a month. I have further discussed about the result in the analysis section. In the future work portion, I have shown my thoughts about the further improvements and limitations of my current methods. My project has only demonstrated an ordinary process for the application of NLP in the investment aspect. With current level of technology such as Artificial Intelligent, the utilization of NLP is highly matured in this field. The project is a great start for me to further research in this specific area.

2 Data Collection and Processing

2.1 Data Collection

For the project, I acquired three datasets related to Tesla stock data, Tesla-related financial news data from Real-Time Finance and Seeking Alpha through APIs and Telsa corporate informations from Wikipedia, respectively. I utilized requests library with proper API keys to finish data collection processes. For data from Wikipedia, I used Beautiful Soup to achieve web scraping. Links of my data sources are provided below:
https://rapidapi.com/apidojo/api/seeking-alpha
https://rapidapi.com/letscrape-6bRBa3QguO5/api/real-time-finance-data
https://en.wikipedia.org/wiki/Tesla,_Inc.#Finances

My original plan was to utilize a Python library created by Finviz website, which was used to acquire financial data by importing packages from the library. However, it cannot demonstrate a solid data scrapping skill for the purpose of the project even though it provides consistent and accurate data. Therefore, I have switched to collect data via APIs. Moreover, I was inspired by Professor Satyukov to collect news article data from LexisNexis website, however, it was not feasible due to restrictions of my current knowledge and a matter of time. I had stayed with news headlines as my dataset.

2.2 Data Processing

After corresponding datasets collected and saved as JSON files in “raw” folder, I processed them by observing and interpreting the nested structure then extracting useful data from the files,

respectively. From "raw_news_data", I extracted "id", "title" and "data_time" for each attribute. In the other dataset, I extracted "date_time", "price" and "volume". Then I both transformed them into Pandas data frames. As I analyzed them as timeseries data and my later processes required the date and time data points in proper datetime format, I aligned the data formats by using "pd.to_datetime". I deleted duplicate datapoints. Additionally, when addressing missing data in the news dataset, I initially filled in the gaps to ensure continuity of dates and filled in NaN for the title of the news on that day, leading to a zero for the sentiment score. For the missing data in stock data, I used forward filling method to fill in missing stock price data points since there were a few days on which stock market was closed. For the data from Wikipedia, I removed the attributes included "zero" and NaN. Then I ensured the format of the data as integers or floats. After processing the data, the next section will explain my analysis methods and data visualization process.

3 Analysis and Visualization

3.1 Analysis Techniques and Related Findings

I applied two analysis techniques for my project. The first method is the Natural Language Processing and the other one is Granger Causality Test. Through the first analysis phase, I extracted sentiment scores from the titles of the news data by using nltk vader library as the the NLP model. The model provided a dictionary including sentiment scores implied by different words. After running the model with the news data, it resulted to a new dataset with sentiment scores related to each news headline of the news data. Next, I conducted the second phase which was Granger Causality test, finding the relationship between the sentiment scores and the stock price in a daily frequency. Figure 1 below has shown the result of the statistical test.

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.4695  , p=0.4989  , df_denom=28, df_num=1
ssr based chi2 test:   chi2=0.5198  , p=0.4709  , df=1
likelihood ratio test: chi2=0.5155  , p=0.4728  , df=1
parameter F test:         F=0.4695  , p=0.4989  , df_denom=28, df_num=1
```

Figure 1. Result of Granger Causality Test

The results from the Granger Causality test indicate that there is no statistical evidence to suggest that one time series can predict another. This conclusion is drawn from the high p-values (all above 0.47), which are much greater than the common alpha level of 0.05. With such high p-values, we fail to reject the null hypothesis of no Granger-causality. The number of lags tested is one, meaning the test checked for causality with a one-period lag. The F-statistic and chi-square values are associated with this test, but they are not significant given the p-values. In summary, there is no evidence proving that the sentiment scores of the TSLA news headlines is associated with the stock price of TSLA.

3.2 Data Visualizations and Explanations

I provided four figures in terms of the data visualization purpose for this project, which better explained different steps through the projects. Figure 2 attached below presented Tesla stock price trend over a year. It showed that Tesla stock price peaked at the end of July in this year over the period and the lowest point was around January 1st, 2023. The overall trend was increasing over the time. The figure was formed by scatter plotting and line plotting. I plotted them using seaborn and matplotlib libraries, respectively.

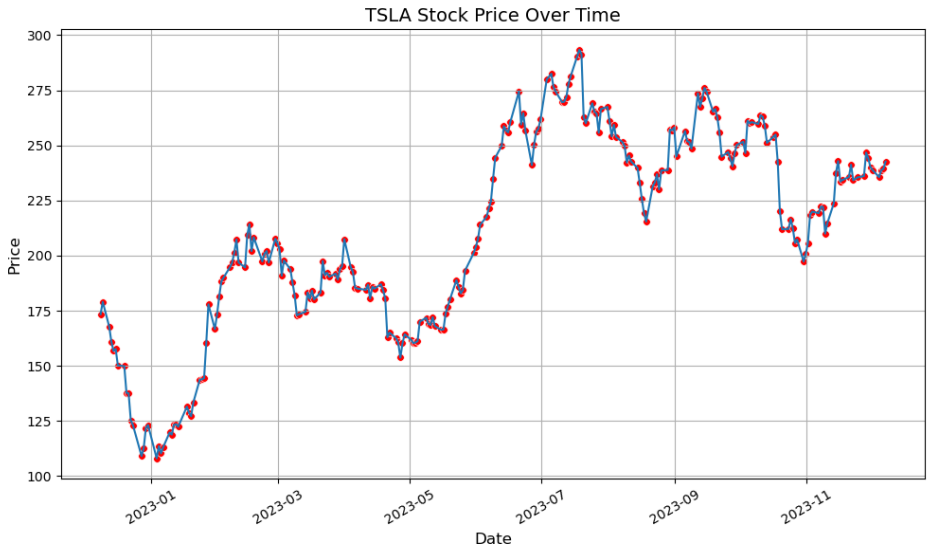

Figure 2. TSLA Stock Price Trend in a year

The second figure (Figure 3) shown below shows a bar chart describing the daily sentiment scores extracted from the news headlines. The chart was created by matplotlib.

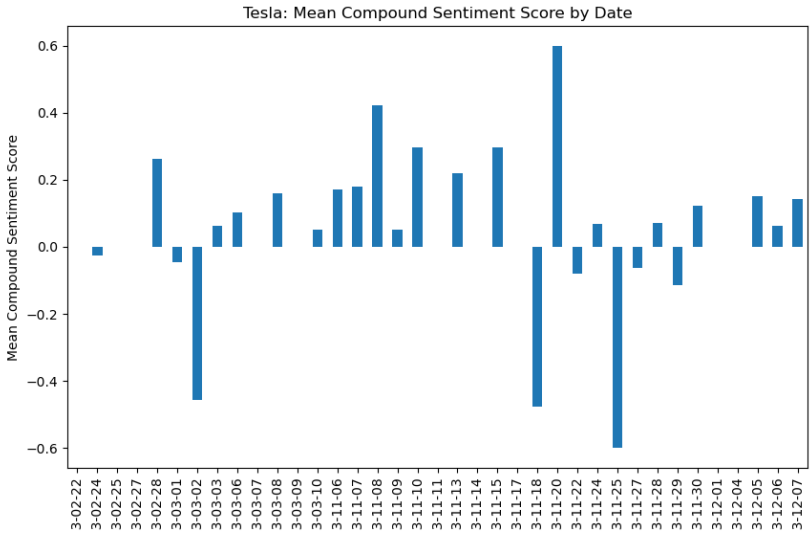

Figure 3. Compound Sentiment Scores by Date

Figure 4 shown below indicates positive, neutral and negative sentiment scores in a daily basis using stacked bar, which was generated with matplotlib library. Each stacked bar is consisted of three portions in green, orange and red, which represents positive, neutral and negative scores, respectively.

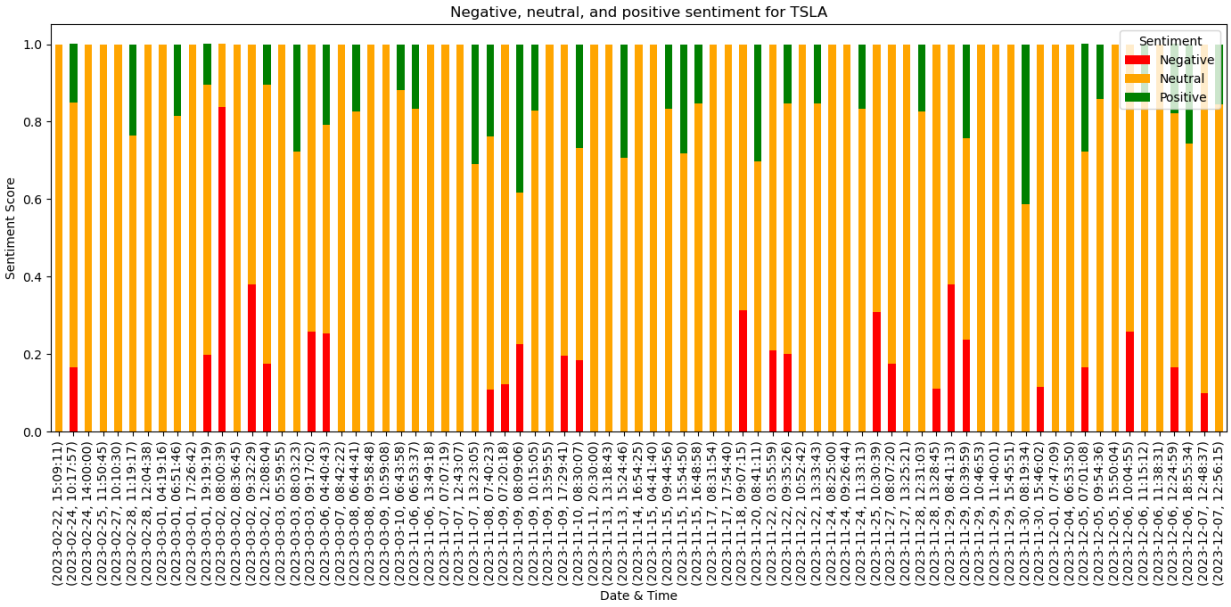

Figure 4. Stacked Bar Chart showing Sentiment Scores

The final figure (Figure 5) presents a comparison of trends of Compound Sentiment Score and Stock Price over a month. The blue line chart shows the Tesla stock trend, and the orange scatter plot shows the compound sentiment scores, both created with matplotlib library.

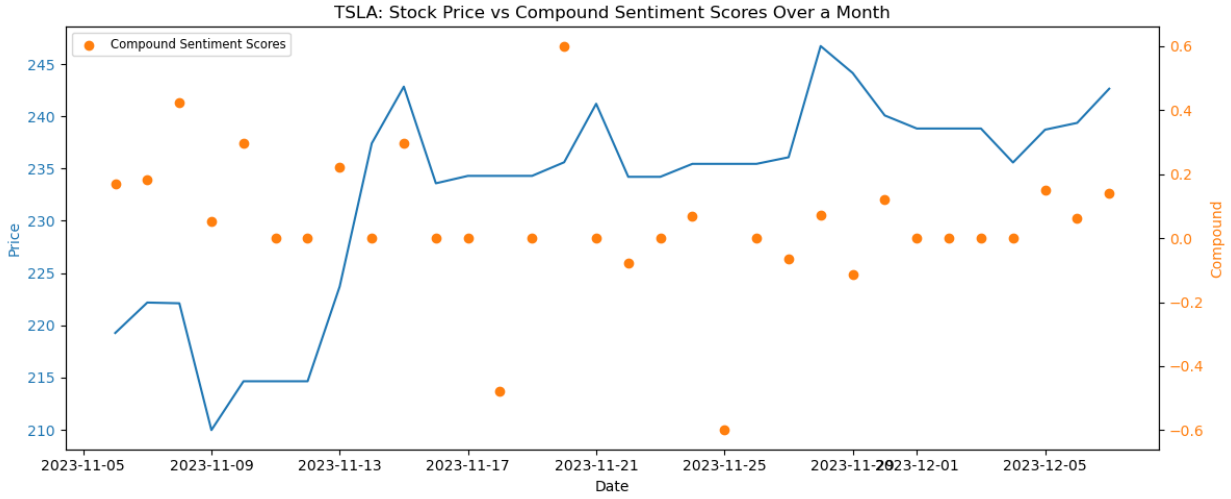

Figure 5. Comparison between Compound Sentiment Score and Stock Price over a month

Figure 6 below demonstrates financial information and employee numbers from 2007 to 2022. A combined graph provides a more comprehensive visualization for multiple sets of data and it is beneficial for a quick comparison when scheming it.

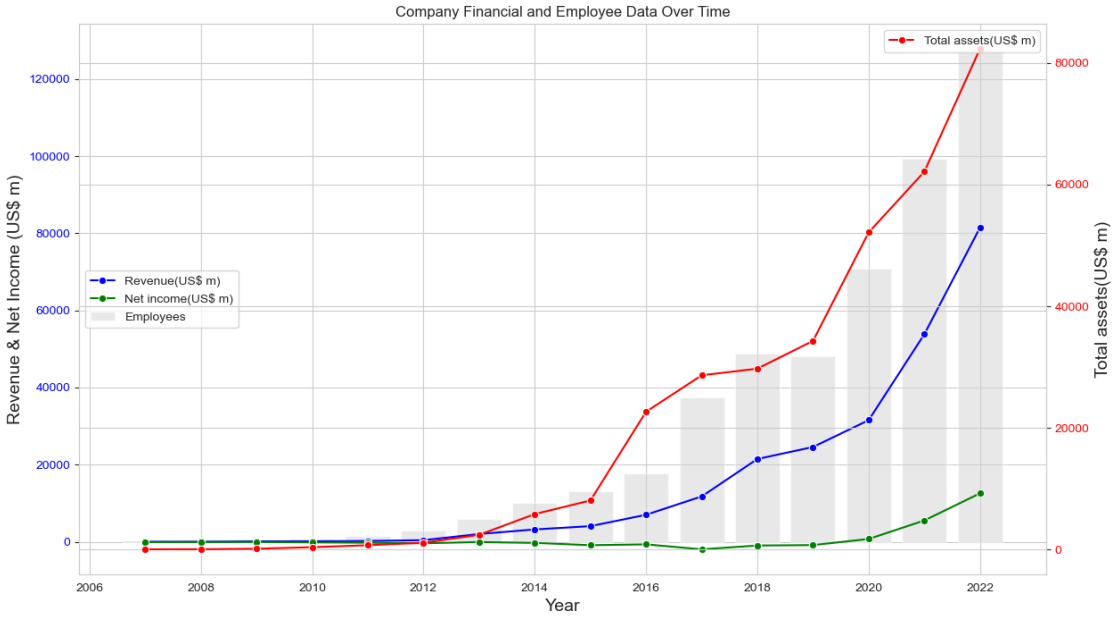

Figure 6. Combined graph Showing Tesla's Financial data and Employees number over Years

The first three figures are for the purpose of visualizing the three datasets. The fourth figure shows my observed conclusion which is that the stock increased when there was related good news in most of the times but when there was negative news, the stock price did not change correspondingly. This also explains the statistical conclusion which is that there is significant association between the sentiment scores and stock price trend in this case. In terms of the impacts of my findings, it reasonably explains that it is worthy for further exploration in this field because there are a few cues leading to association. The final figure provides a comprehensive overview of Tesla's corporate information, showcasing financial trends and employee numbers over the years.

4 Future Work

Given the project's limitations, future work could include:

1. Analyzing full articles rather than just headlines to potentially increase the accuracy of sentiment scores, as headlines may not fully capture the article's tone.
2. Employing a more sophisticated NLP model with a finance-specific lexicon could provide more accurate sentiment analysis for this research.
3. In the analysis, using machine learning models, particularly pre-trained transformer models for NLP, might be a more suitable approach than the dictionary method. These were not utilized in this project due to my current limitations in machine learning expertise.
4. An interactive dashboard would be a superior choice for data visualization, offering the capability to adjust the time period and select different companies.