Final Project Report

Group 6: Cristian Leo, Liying Liu, Pataris Chaipromprasith,

Po Wen Hsu, Raja Navaneeth Talluri

Applied Analytics, Columbia University

APAN 5205: Applied Analytics Frameworks & Methods II

Birol Emir

May 18, 2023

Description of Research Problem

In stock market prediction, fundamental and technical analysis are the two traditional methods used to predict the stock markets. Fundamental analysis involves analyzing the security's intrinsic value (related economic and financial factors). The intrinsic value reflects the company's financial and macroeconomic conditions. (Segal, 2022) The fundamental analysis will determine whether the stock price is undervalued or overvalued by the investors. The technical analysis predicts the stock price movement by analyzing the historical data, patterns, and technical indicators. However, with the emergence of the internet and social media, sentiment analysis has been developed to predict stock prices by analyzing people's posted opinions. "Sentiment analysis is a field of study that deals with the people's concerns, beliefs, emotions, perceptions, and sentiments towards some entity." (Rouf & Somanathan, 2021) Since a rising stock price is due to investor confidence, sentiment analysis could lead to more accurate predictions. (Park, 2022) Many platforms are well known for extracting sentiment for investors to analyze, such as Yahoo Finance, Twitter, The Stock Twit, etc. Given this, we have decided to analyze the Twitter post relating to Tesla to see if sentiment affects the stock price.

Literature Review

The stock market prediction has always been a popular subject for scientific researchers and finance professionals. Thanks to the progress of technology, everyone with the internet can easily approach financial reports and search articles about a particular company's financial situation. Moreover, people can influence stock market performance to some degree. In recent years, social media has become an increasingly important platform for people to exchange opinions or update information. Research has shown that social media can predict stock market performance (He, Guo, Shen, and Akula, 2016). Take the GameStop frenzy in January 2021 as an example. The short was countered by active members of the social media group wallstreetbets on Reddit.com (Yang 2022). Similarly, another study illustrates that social media and the expansion of information sources have affected the stock market price and must be considered to improve the accuracy of stock market prediction (Ishikawa, Sassi, and Yahia, 2021). However, the study also mentioned the problem of fake and misleading information from social media.

Compared to traditional research methods, such as surveys and polls, social media has the advantage of people's exact and instant sentiments. Since sentiment can highly reflect people's thoughts, extracting emotions and opinions from social media platforms is another direction to analyze data. Many studies are on developing data mining and natural language processing (NLP) techniques to seek the relationship between social media and stock market price. One study focused on Twitter data and constructed a Tweet Node Model to make a prediction (Ni, Wang, and Cheng, 2021). Another study developed a novel stock movement predictive network model

(SMPN), combining Tweet and historical stock market price observation (Xu, Chai, Luo, and Li, 2020).

Generally, there are three advantages of sentiment analysis. First, it is more cost and time efficient compared to traditional methods like surveys. Second, it extracts people's opinions in real-time, avoiding recall biases. Third, it can provide a temporal sentiment profile (Philander, and Zhong, 2016). Sentiment analysis provides us with another aspect of understanding the data.

Overall, the above literature review triggered our curiosity about how social media and users comments will influence the stock market. However, due to the limited time, we only selected one key opinion leader (KOL), Elon Musk, the founder of Tesla, who is actively involved in Tweet, as our approach to analyzing this phenomenon.

Research Question To Be Examined

Given the background, the research question that will be examined is "Based on Elon's Twitter post relating to Tesla, does Elon's post has an effect on Tesla price and how well does it help with predicting the fluctuation (increase or decrease) of the stock price?" The research question will examine whether the sentimental value of people on Twitter affects the stock price and if it correlates with the price. The result of the study will provide evidence of whether the sentimental value from Twitter's Tesla post should be used for predicting stock price.

Data Description

Data Source: https://www.kaggle.com/datasets/vidyapb/elon-musk-tweets-2015-to-2020

Ideally, Twitter API can be utilized to extract Elon Musk's tweet from Twitter. However, Twitter API is not accessible to everyone and requires users to apply for access. Given this, the dataset from Kaggle was utilized to conduct the analysis. The dataset includes date and tweet columns, as well as several columns containing the Corpus extracted from Elon Musk's tweets. The Corpus contains single words extracted from the tweets, with each single word being assigned to a specific column after vectorization. In addition, bigrams and trigrams were generated from the cleaned text data to create additional features, which were assigned to every single column.

The dataset also includes columns containing the word scores, which provide information on the sentiment of each tweet, categorized as positive, negative, or neutral. Moreover, the dataset includes Tesla stock price data downloaded from Yahoo Finance, which was merged with the

original dataset to enable analysis of the relationship between the sentiment of Elon Musk's tweets and Tesla stock prices over time.

Overall, this dataset is designed to provide insights into the sentiment of Elon Musk's tweets over time and its impact on Tesla stock prices, with additional features extracted from the text data to help identify trends and patterns that can be useful for generating insights and making informed decisions.

Reasons behind the choice of analytical techniques

In this project, we aimed to predict whether the Tesla stock price would go up or down based on Elon Musk's tweets. We used R to preprocess the data, perform exploratory analysis, engineer features, and model the data.

Firstly, we preprocessed the data by grouping tweets by date, creating a corpus, converting text to lower case, removing URLs, punctuation, stopwords, and whitespaces, lemmatizing, removing non-alpha characters, vectorizing, and performing sentiment analysis using Vader. We also used the Yahoo Finance API to merge the Tesla stock price data with our text data.

Next, we conducted exploratory analysis to gain insights into the data. We analyzed the counts of increases and decreases, top 25 words, sentiment score over time, top positive and negative words, emoji correlation with stock price, and average stock price per emoji. This analysis allowed us to identify trends and patterns in the data, which were essential for creating accurate models.

After exploring the data, we engineered features to improve the performance of our models. We created 2 n-grams, 3 n-grams, and filled in missing data.

Finally, we modeled the data using a deep learning model in Keras. The model architecture included two LSTM layers and a dense layer, and it was trained using binary cross-entropy loss and the Adam optimizer. The model was also evaluated using accuracy as a metric. To create the model, we first determined the number of time steps and features in the time series data and the number of words and dimensions in the NLP data. We then set the number of LSTM units and dense units for the model, reshaped the time series and NLP data to match the input shape of the model, defined the input shapes for the time series and NLP data, and created the model architecture. We compiled the model using the compile() function and defined a callback to save the best model based on validation accuracy. We trained the model using the fit() function,

including options for the number of epochs, batch size, validation split, and verbose level, and evaluated the model using the evaluate() function.

In conclusion, the prediction of stock prices based on social media data is a challenging but rewarding task in the field of data science. Our project demonstrated the importance of preprocessing, exploratory analysis, feature engineering, and modeling to create accurate and robust models. By combining these techniques, we were able to successfully predict whether the Tesla stock price would go up or down based on Elon Musk's tweets.

Analysis Steps and Process

- Collection of Data:

In an attempt to get fresh data directly from source, we experimented with APIs of Twitter. However, we could not get the approval in time for the access to the API. Therefore, we settled on using already scraped data from Elon Musk's official account from 2015 to 2020 on Kaggle. Adjusted closing stock price along with the date of stock price is extracted from Yahoo Finance. The stock price data time range matches with tweets data time frame (2015 to 2020).

- Data Cleaning:

The dataset consists of 9286 tweets along with 33 other variables related to the tweet like date and time of the tweet, id of the tweet etc. Of all the 34 variables we only need the tweet and date of the tweet for our analysis. We filtered out the rest of the columns from the source dataset. Next tweets are grouped by date by concatenating the tweets tweeted on a single day and arranged in descending order. The latest tweets appear first in the table. To perform various text analysis tasks such as cleaning and preprocessing the text, creating a document-term matrix, and conducting text mining operations like topic modeling or sentiment analysis, a corpus object is created. A corpus object is a collection of text documents that have been processed and organized for text analysis. Once the corpus is created, we perform standardization operations across all the tweets. To achieve uniformity with respect to case, we convert all letters of the alphabet to lowercase. We removed certain parts of the tweet's content that does not affect the stock price like embedded URLs, stop words, punctuations and whitespaces.

- Feature Creation:

We perform lemmatization to the corpus of documents by extracting the lemma columns from tokenized documents, removing non-alphabetic characters, and returning the lemmas. From the preprocessed and cleaned lemmatized documents, we create a document term matrix. The DTM is normalized by dividing each count by the total number of words in the document. The DTM is

converted to a matrix and any term that appears in less than 5 percent of the documents is removed. We then compute the TF-IDF values for each term in the DTM by sorting the column sums of the data frame in descending order. This gives the most frequent terms in the corpus, weighted by their inverse document frequency. Next we analyze the sentiments of a collection of tweets and calculate a sentiment score for every tweet. We create a new data frame with six columns, one for each type of sentiment score calculated: word_scores, compound, pos, neu, neg, and but_count. We also create new features from tweets using 2 grams and 3 grams technique. To finalize the predictor variables, we merge the TF-IDF, sentiment scores and n-grams data frames with the original data frame containing tweets and date of tweet.

- Unified Table Creation:

We join the input variables data frame with the stock price data frame on the date column to get all the tweets and the corresponding stock price on a particular day, which is the predicted variable.

- Model Building and Prediction:

The process of building the model and making predictions involves the use of neural networks to model the relationship between input variables and stock prices. First, the data is split into training and validation sets, with 80% used for training and the remaining 20% for validation. Two types of data are used in the model: time series data, which comprises only the close adjusted stock price column of the dataset, and NLP data, which includes the TF-IDF matrix, Vader sentiment scores, and n-grams. Two separate neural networks are used for each type of data: a LSTM neural network with one hidden layer of 32 nodes and a ReLu activation function is used for the time series data, while a normal neural network with one hidden layer of 64 nodes and a ReLu activation function is used for the NLP data. The outputs of these neural networks are then fed as inputs to a new neural network with a single output, which is either 0 or 1. The final neural network uses a sigmoid activation function, and the model is compiled using Adam optimization, trained over 2000 epochs with a learning rate of 0.01. The model's hyperparameters, such as hidden layers, nodes, activation functions, learning rates, and epochs, were experimented with, and the aforementioned values were found to maximize accuracy.

The model architecture for the advanced version consists of two LSTM layers, one for time series data and one for NLP data, followed by dropout layers to prevent overfitting, dense layers with rectified linear unit (ReLU) activation function, and a concatenation layer to merge the outputs from the two LSTMs. Finally, a dense layer with sigmoid activation function is used to obtain the binary classification output. To make predictions, the model is first trained using the fit() function, which trains the model on the input data and labels. Options such as the number of epochs, batch size, validation split, and verbose level can be specified. The callbacks defined earlier for early stopping and saving the best model are also included. For the evaluation of the model, the evaluate() function is used, which computes the loss and accuracy of the model on the test data. The load_model_hdf5() function is used to load a saved Keras model from an HDF5 file, and the

evaluate() function is used to evaluate the model's performance on the test data. The input data and labels are provided as lists, and the test loss and accuracy are printed.

- Model Results:

In the first model, we achieve 75% test accuracy to predict the stock price fluctuations based on Elon Musk's Tweets. The second model performance result is 54% test accuracy. This indicates that the first and simpler model performs better than the second and more complex one. The results of running the code are as follows:

Model 1:

```
# Evaluate the model
scores <- model %>% evaluate(
  x = list(time_series, nlp_data),
  y = labels,
  batch_size = 32
)

# Print the evaluation metrics
cat("Test loss:", scores[[1]], "\n")
```

```
## Test loss: 0.9218721
```

```
cat("Test accuracy:", scores[[2]], "\n")
```

```
## Test accuracy: 0.7489083
```

Model 2:

```
# Load the best saved model
library('keras')
model <- load_model_hdf5("best_model_advanced.h5")

# Evaluate the model on the test data
test_loss_and_metrics <- model %>% evaluate(
  x = list(time_series_test, nlp_data_test),
  y = test_labels,
  batch_size = 32
)

# Print the test loss and accuracy
cat(paste0("Test loss: ", test_loss_and_metrics[1], "\n"))
```

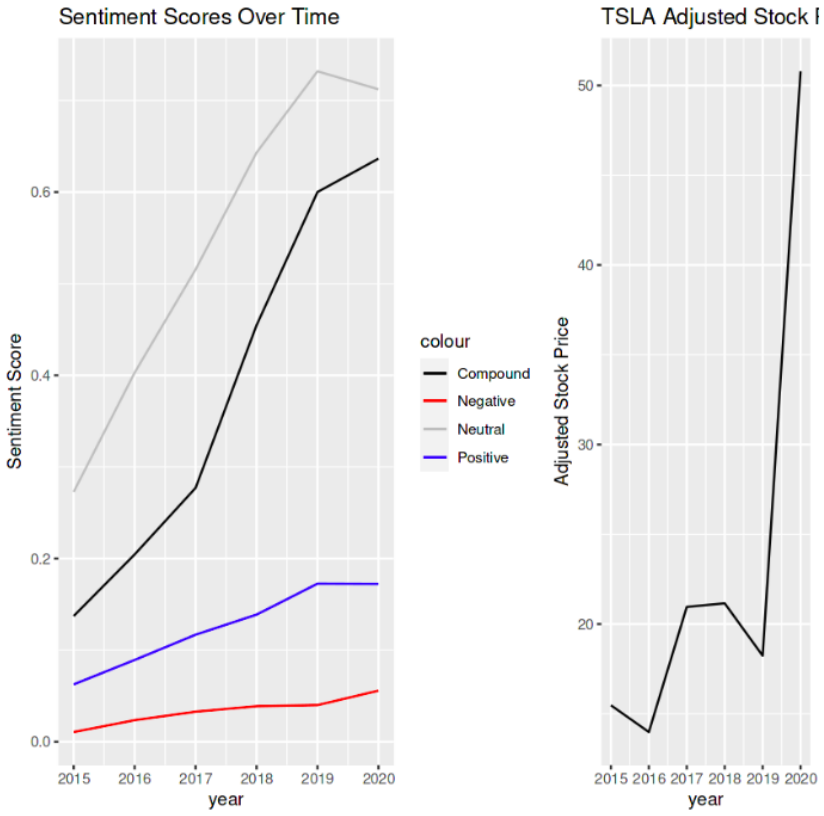
```
## Test loss: 0.704934179782867
```

```
cat(paste0("Test accuracy: ", test_loss_and_metrics[2], "\n"))
```

```
## Test accuracy: 0.538181841373444
```

Conclusion and recommendations

Based on our analysis effort, we obtained two conclusions. First, the relationship between tweets and stock fluctuations is low. From the graph, we can see that the trend of positive and negative sentiment scores over time did not match the trend of Tesla's stock price. On the other hand, the compound and neutral sentiment scores have a more similar trend to Tesla stock. Hence, the correlation provides evidence that Elon Musk's Tweets may have some impact on Tesla's stock price regardless of the content (positive or negative).



Secondly, the more advanced model may need more complex feature engineering and feature selection from a dataset. We ran two models to verify the second part of our research question "Based on Elon's Twitter post relating to Tesla, does Elon's post has an effect on Tesla price and how well does it help with predicting the fluctuation (increase or decrease) of the stock price?". Our simple model got 0.7489 out of 1, and our advanced model got 0.5381 out of 1. The reason why we got the result may be that we used too many features in a prediction model, which can lead to a problem known as "overfitting." Overfitting occurs when a model becomes too complex and starts to fit the noise or random fluctuations in the data rather than the underlying patterns. To avoid overfitting, it is important to carefully select the most relevant features to solve the problem and use techniques such as regularization and cross-validation to prevent the model from becoming too complex. Moreover, the advanced model split the data base on the index time series in the

analysis. Given this, splitting the data through time series could lead to lower accuracy and it might provide evidence that time may not have much effect on the trajectory of the stock price.

Sentiment analysis can provide valuable insights of public perception of a particular topic or brand. However, it is important to keep in mind that sentiment analysis is not a perfect science, and its accuracy can be influenced by many factors, such as the language used in the tweet, the context in which it was made, and the culture and demographics of the people responding to it.

When it comes to the stock market, many variables can influence stock prices, such as financial reports, economic indicators, and industry trends. While sentiment analysis can be a helpful tool in predicting the impact of certain events or a specific person on a stock's performance, it should not be the only factor considered. Since our prediction model only included Elon Musk's Tweets. It means it will be precarious to make decisions by relying on our prediction model or any other model that only considers some factors.

Take our analysis as an example. The impact of a single tweet on Tesla's stock price can be challenging to predict. Even if the tweet's sentiment is positive, other factors, such as the current state of the market or the company's financial performance, can still negatively impact the stock price. Conversely, a negative tweet may not necessarily result in a decline in stock price, as other factors could offset the impact.

Therefore, it is essential to approach investment decisions cautiously and not rely solely on sentiment analysis. Conducting thorough research on a company's financial performance, market trends, and industry analysis such as SWOT, PESTLE, and Porter 5 Force analysis can provide a more concrete picture of its potential for growth or decline.

Consulting with a financial specialist or conducting additional research can help to mitigate risks and make informed investment decisions. Finally, it is important to remember that investing in the stock market always carries some risk. It is crucial to have a diversified portfolio to minimize potential losses.

Work Cited

He, W., Guo, L., Shen, J., & Akula, V. (2016, April). Social Media-Based Forecasting: A Case Study of Tweets and Stock Prices in the Financial Services Industry. Columbia University Libraries. Retrieved February 27, 2023, from https://go-gale-com.ezproxy.cul.columbia.edu/ps/i.do?p=ITOF&u=columbiau&id=GALE%7CA447828213&v=2.1&it=r&sid=summon

Ishikawa, T., Sassi, I. B., & Yahia, S. B. (2021, May 8). Assessment of Malicious Tweets Impact on Stock Market Prices. Columbia University Libraries. Retrieved February 27, 2023, from https://link-springer-com.ezproxy.cul.columbia.edu/chapter/10.1007/978-3-030-75018-3_22

Ni, H., Wang, S., & Cheng, P. (2021, April 22). A hybrid approach for stock trend prediction based on tweets embedding and historical prices. Columbia University Libraries. Retrieved February 27, 2023, from https://link-springer-com.ezproxy.cul.columbia.edu/article/10.1007/s11280-021-00880-9

Park, Meeyoon. (2022, February 22).Market Sentiment.Corporate FInance Insititute .https://corporatefinanceinstitute.com/resources/capital-markets/market-sentiment/

Philander, K., & Zhong, Y. Y. (2016, April). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. Columbia University Libraries. Retrieved February 28, 2023, from https://www-sciencedirect-com.ezproxy.cul.columbia.edu/science/article/pii/S027843191630007X?via%3Dihub

Rouf, Nusrat & Somanathan, Arjun (2021, December 9). Stock Market Prediction. Scholarly Community Encyclopedia. https://encyclopedia.pub/entry/16165

Segal, Troy. (2022, August 25). Fundamental Analysis: Principles, Types, and How to Use It.Investopedia.https://www.investopedia.com/terms/f/fundamentalanalysis.asp#:~:text=Fundamental%20analysis%20is%20a%20method,a%20buy%20recommendation%20is%20given.

Usmani, S., & Shamsi, J. A. (2021, May 4). News sensitive stock market prediction: Literature review and suggestions. PeerJ. Computer science. Retrieved February 27, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8114814/

Xu, H., Chai, L., Luo, Z., & Li, S. (2020). Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices. Columbia University Libraries. https://www-sciencedirect-com.ezproxy.cul.columbia.edu/science/article/pii/S0925231220313060?via%3Dihub

Yang, Z. (2022, December 27). GameStop or game just started? leveling the playing field for social media meme investors to rebuild the public's trust. MDPI. Retrieved February 27, 2023, from https://www.mdpi.com/1911-8074/16/1/13