**Stock Market Analysis informed by CEO Social Media Communications**

**Efsa Sema AKTAY**

1. **Introduction**
   1. **Background**

Traditionally, the companies have been evaluated for their fair pricing through fundamental analysis. Fundamental analysis utilises multiple factors by evaluating corporate performance, interest rates, and discount rates to forecast to free cash flow and arrive at a fair value price. Although, the financial markets have been generally influenced by a variety of factors, including economic indicators, geopolitical events, and corporate performance; in recent years, social media has emerged as a force capable of changing market sentiment and influencing the movement of stock prices. Social media communication especially from influential CEOs, have shown to have an impact on market behaviour. To exemplify, the instance such as Elon Musk's tweets about Tesla, provide a context for exploring this phenomenon. Specifically as seen in the Figure 1 below, Elon Musk's infamous tweet about taking Tesla private at $420 per share led to massive market reaction and SEC scrutiny.

A graph with a red arrow pointing to the top

Description automatically generated

*Figure 1: Displaying the stock price decrease following Elon Musk’s tweet.*

As seen on Figure 2 below, another significant shock on the market is the case of GameStop stock price had received significant increase in start of 2021 through the Reddit community called r/wallstreetbets through the simultaneous purchase of the stock. High-profile cases, such as the Tesla and GameStop stock surges, highlight the significant impact of social media activity on stock prices. (Wang et al., 2024) For instance, the GameStop short squeeze in early 2021 was heavily influenced by discussions and coordination on the subreddit r/wallstreetbets, illustrating the power of collective social media behaviour in driving market dynamics.

A graph of a price fall

Description automatically generated with medium confidence

*Figure 2: Displaying the stock price increase in the case of GameStop.*

Despite the growing influence of social media communications, the quantitative impact of CEO tweets on market movement have not been extensively researched and documented. Specifically, the mechanism for which these communications affect stock market volatility and price movements provide an opportunity for further research. The objectives include firstly, gauging whether there is a quantifiable impact of CEO tweets on stock market volatility and price movements of the stocks. Secondly, developing models that can efficiently utilise the text input within the predictive models for stock market forecasting.

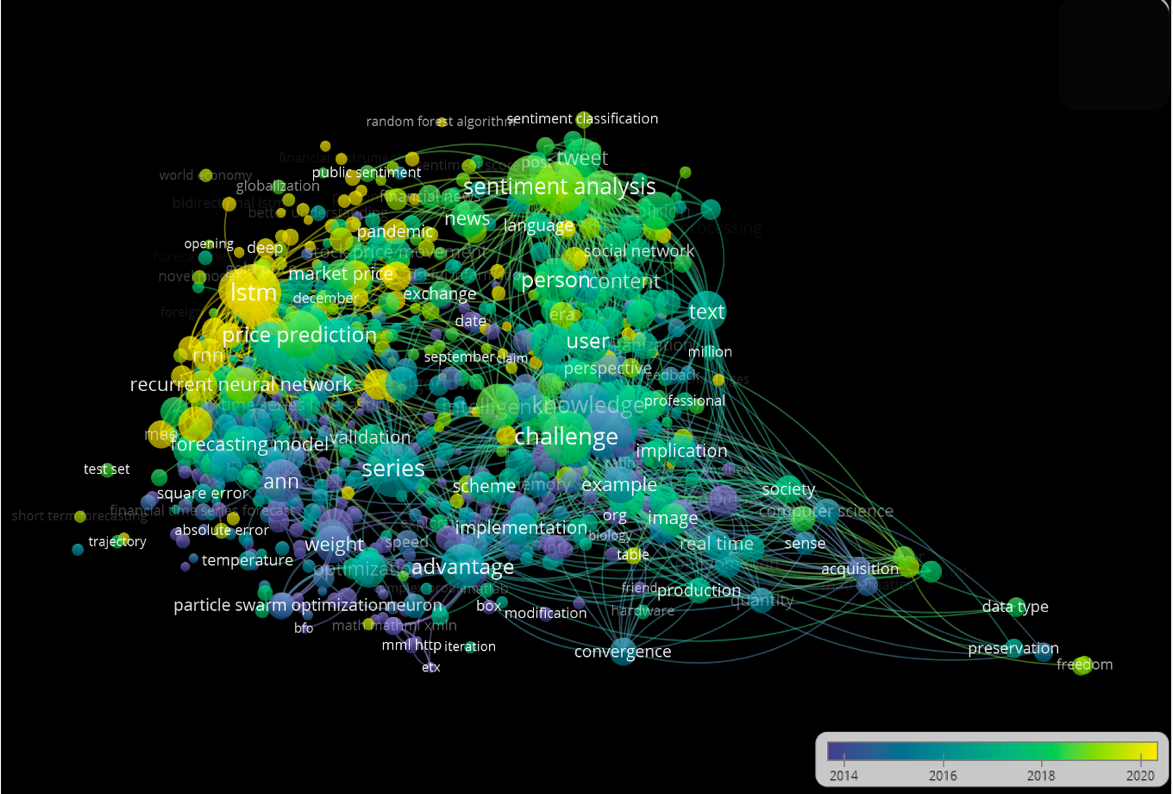
Exploration of the influence of non-traditional market, such as social media communication can help enhance the accuracy of predictive financial models, provide windows of opportunity for profit or help reduce the downside risk for the stakeholders. This analysis could provide stakeholders such as investors, financial analysts and managers with new tools and insights for market analysis.

1. **Literature Review & Theories**
   1. **Efficient Market Hypothesis and Random Walk Theory**

There are two major theories that make up the foundation of the stock forecasting. Firstly, the random walk theory suggests that stock price changes are random and do not follow a predictable pattern. Secondly, the efficient market hypothesis(EMH), proposed by Fama (1970) reflects that stock prices reflect all the available information, proving it quite impossible to generate returns that are consistently higher than the overall market. According to EMH, any new information is quickly incorporated into stock prices, rendering them unpredictable. The study is exploring the opportunity for discerning the market sentiment through social media information analysis and recognises the randomness of the stock market data.

**B. Literature Review**

Existing sentiment analysis techniques have been applied to financial predictions with varying degrees of success. (Todd et al., 2024) Previous studies have employed various machine learning models, including linear regression, support vector machines, and neural networks, to predict stock price movements based on sentiment analysis. As visible in the Figure 3, a cluster algorithm displaying the recent literature with the cluster algorithm representation. It is possible to see how LSTM and sentiment analysis have been more relevant after 2022. This is the result of the recent development in the machine learning and LLM technology improvement in the recent years. The cluster algorithm has been generated according to the density of the interconnectedness of the research in stock market prediction.



*Figure 3: Displaying the existing research in a cluster algorithm.*

In addition, the cluster algorithm and the connecting lines have been generated according to the strength of their connections with other perspectives on the topic. For example, if there are a lot of research connecting tweet data and sentiment analysis for stock

A colorful network of dots and lines

Description automatically generated with medium confidence

*Figure 4: Displaying the connectedness of the recent research concepts in a cluster algorithm.*

market prediction both of these concepts will be highlighted and displayed. As seen in the Figure 4 utilising sentiment analysis in combination with tweet data is as largely explored and document as other traditional stock market prediction models as of this moment.

1. **Methodology**

A diagram of a data processing process

Description automatically generated

*Figure 5: Displaying the flowchart of the overall process*

In the figure above the methodology has been displayed in a flowchart, the steps of the modelling and the process will be explored in the following sections.

* 1. **Data Collection and Generation**

The data that is utilised in the model can be grouped into two, the synthetic data that is generated and the real company and CEO tweet data. The synthetic data generation process was informed by the real life scenarios in some aspects of the tweeting and the stock behaviour. For instance, tweeting frequency, how impactful the tweet is to the stock price and the correlation between the stock price and the index. This synthetic data experimentation will enable a more controlled environment for the model and the impact of tweets to be analysed as a base case.

***Synthetic Data***

The synthetic data was created by the help of GENAI specifically, ChatGPT-4 to simulate the tweeting patterns and corresponding profiles of companies. The overall process can be viewed from the Figure 6 below.

A diagram of a data processing process

Description automatically generated  
*Figure 6: Displaying the flowchart of the synthetic data generation process*

The algorithm generated tweets, stock, and index prices. The entire dataset were generated over the period 2020-01-01 to 2024-04-15 timeline, reflecting the number of dates in the real data. By assigning the random seed number it was ensured that the stock, index and tweet data can be reproduced for verification purposes. Firstly the tweet generation process will be explored. The tweet generation algorithm is set up so that there are days without tweets and days with multiple tweets which is an observed behaviour in our real life CEO’s twitter accounts. Then, three imaginary CEO’s were initialised to represent a profile of CEO’s in real life. Overall, CEO1 represented a CEO with a frequent and manipulative tweeting behaviour tweeting many positive and negative sentiment tweets. CEO2 represents a more positive and less manipulative tweeting behaviour. CEO3 represents a more neutral baseline case with the least frequent tweeting behaviour and the least manipulative tweets. As seen above in the Figure 6 the CEO’s were scored out of 10 for the sentiment of their tweets which enabled the tweet generation process. CEO1 is assigned the most negative with 70% negative 15% positive and 15% neutral sentiment tweets as well as tweeting the most frequently. CEO2 is assigned 15% negative 70% positive and 15% neutral sentiment tweets with a moderate tweeting frequency and the CO3 was assigned 15% positive, 15% negative and 70% neutral sentiment tweets and a relatively modest tweeting frequency. In the Figure 7 below, some sample tweets for each sentiment is displayed in a table for reference. The red underlined words are some of the sample words that is swapped into different sentence structures to generate these tweets.

The negative cases tend to feature information regarding sales, supply chain issues, and competitive pressures; the positive cases are about the launch of new products, features, and record sales; and the neutral cases are mostly about corporate announcements, earnings calls, and meetings.

**A close-up of a message

Description automatically generated**

*Figure 7: Displaying the sample sentences from the synthetic data generation process*

In addition the tweeting frequency and the sentiment of the tweets for each CEO has been visualised for better comprehension below at Figure 8. For the purposes of the below visualisation the sentiment is scored with the following criteria: negative, positive and neutral sentiment is scored as -1, 1 and 0. It is possible to see that line chart is very dense for CEO one with a concentration on the negative sentiment. For COE2, tweeting is less dense and mostly positive and for CEO3, it is the least frequent with mostly neutral tweets. The visualisation below validates and visualises the systematic tweet generation process.

A screenshot of a graph

Description automatically generated

*Figure 8: Chart displaying the tweeting frequency and the sentiment in the synthetic data*

Following the generation of the tweets, the sentiments of each tweet has been stored in a data frame. Three different stock and index prices were generated systematically namely Stock 1, Stock 2, Stock 3, Index 1, Index 2 and Index 3. In the generation method, the stocks had a randomised correlation with the tweet. For example, the sentiment data frame were utilised to initialise a change in stock price between 0-10% increase for positive tweets and or 0-10% decrease in the stock price. The indexes were generated separately and randomly. However, in addition to the tweet data, a random correlation with the index price movement has also been introduced to the stock. This generation method is informed by both the real life cases and the random walk theory, complementing both financial theory and preparing a comparable model for analysis. Below in Figure 9, the stock data has been first normalised and plotted together with the index prices for inspection purposes.

A graph of stock indexes

Description automatically generated

*Figure 9: Visualisation of the generated Stock1,2and 3 alongside Index1,2,3 plotted after being normalised for clearer representation.*

***Real Tweet, Stock and Index Data***

For the real data, the stock price data of Apple and Tesl,a alongside Nasdaq, S&P500, and Dow Jones Industrial indexes, were retrieved from Yahoo Finance between 2022-01-01 to 2024-04-15. The sample data frame featuring the stock and index data has been included for reference in the Figure 10 below.

A screenshot of a computer

Description automatically generated

*Figure 10: Displaying the scaled data frame of stock price containing tesla, apple and major US stock indexes.*

Below, the Apple and Tesla stock prices were normalised and plotted alongside the three major US indexes in Figure 11 and Figure 12.

A graph of different colored lines

Description automatically generated

*Figure 11: Apple (AAPL) stock data normalised and plotted together with Nasdaq, S&P500 and Dow Jones Industrial Indexes*

**A graph showing the price of a stock market

Description automatically generated**

*Figure 12: Tesla (TSLA) stock data normalised and plotted together with Nasdaq, S&P500 and Dow Jones Industrial Indexes*

As can be viewed from the above normalised plots of the stock data, and an analysis on the tweeting behaviour of Elon Musk and Tim Cook has become a factor in the selection of the companies for the purposes of this investigation. Firstly, Apple is a stable company with a reasonable correlation US market visible through the movement of the index prices. Moreover, CEO Tim Cook’s tweets are largely made up of company announcements and product launches of iPad, iPhone and other products which is similar to the tweeting behaviour of CEO2 and CEO3 in the synthetic data. On the other hand, Tesla is a newer company with a price less correlated with the US market at times the stock price increasing despite a below expectation earning expectation. As a new electric car company, the stock price has been influenced by not only the earning calls and company figures but also the large following of Elon Musk and his social media communication which presents a good contrast between two companies and enables the investigation. The tweeting behaviour of Elon Musk has been largely represented by the CEO1 in the synthetic data. Although, in the synthetic data the generation of political or largely controversial information has been largely refrained. This also exemplifies how the synthetic data generation has been informed by the real life cases for analysis. The stark contrast between the two CEOs tweeting behaviour allows for forming conclusions regarding the market prediction model.

In addition, the corresponding tweets were sourced from X ApiV2(formerly known as Twitter) through the Api subscription. The Basic tier of Api subscription were utilised in this research which introduced certain limitations such as rate limits such that only a certain number of tweets can be scraped from the source, timeline limits limiting the number of tweets that can be scraped from a users’ X account and the overall limits on the scraped tweets is a possible constraint for the comprehensive analysis. Tweets were collected corresponding CEOs which are Elon Musk and Tim Cook between 2022-01-01 to 2024-04-15. Below in Figure 13 a sample data frame has been shown from the tweets of Tim cook. Restrictions placed on the X Api can be noted as a limitation on for the purpose of reproduction of the study.

A white background with black text

Description automatically generated

*Figure 13: Displaying the scraped tweet data frame of Tim Cook.*

Moreover, due to the highly frequent nature of the tweets, the tweets that are posted on the same day has been compiled in one date. For example if a Elon Musk tweeted 6 times from 9 am till 11pm, all the tweet information has been compiled alongside the close price of the stock data in the same date. This methodology introduces another limitation on the study, for further and more comprehensive research utilising real-time stock price as an input and a more robust corporate subscription to X Api with real time tweet pull could help reduce this bias. Another suggestion could also be to introduce high and low stock price that has been seen during the day to monitor the volatility of the stock in a more comprehensive manner.

* 1. **Data Processing**

The data frames has been organised and compiled through the use of pandas library in python. In addition, BERT (Bidirectional Encoder Representations from Transformers) was utilised for processing the text data which was embedded as an input into the deep learning model. The output from the embedder was integrated with stock price data to develop the stock price prediction model. In most of the existing research, the text data is analysed through the sentiment analysis which grouped the text data such as financial data into positive negative and neutral and included scoring on aspects such as polarity and subjectivity. (Todd et al., 2024) Although, the machine learning use is becoming more common after 2022. (Das et al., 2024) In this study the BERT embedder capture complex patterns in text, making them well-suited for analysing tweet data. The embedding transformers the text data into a vectors. An example of the analysis has been presented below. On the left hand side, excerpts from the synthetic data has been displayed on the right hand side the pairwise similarity matrix is presented to visualise how the embedding analysed and vectorised the text data in relation to other text information. This methodology provides a more robust method compared to simple sentiment analysis which is not able to capture the complexity in the text data.

**A diagram of a graph

Description automatically generated with medium confidence**

*Figure 14: Displaying the embedding in a Pairwise Similarity Matrix*

Moreover, a type of deep learning algorithm called Long Short Term Memory (LSTM) was utilised train the model. LSTM is a specialised form of Recurrent Neural Network (RNN) which was selected due to its ability to handle the temporal nature of the timeseries data and dynamics of tweet data. The extensive exploration of the LSTM and BERT architecture is out of scope for this study but both could be explored further to increase the robustness of the analysis.

**C. Validation**

The model was validated using both real and synthetic datasets through two different sets of tests. In both cases, to ensure robustness and reliability of the result, the set of stock and index data is trained both with and without the tweet data and the results are compared. The validation process is completed through comparing the model's predictions with actual stock price movements and evaluating its performance through the root mean square error (RMSE). A limitation of the study is the lack of K-fold validation method which is more representative than a resulting RMSE score to reflect the model’s utility. The study will be further developed to include this in the future analysis.

1. **Analysis**

**Moving Averages**

The predictive model was compared with traditional financial models, such as the Moving Average model which is included in the Appendix below, this method primarily rely on historical stock price data.

**A diagram of a process

Description automatically generated**

*Figure 15: Flowchart displaying the process of Moving Averages Analysis*

According to the model, when the 9 day moving average is above the 21 day moving average the stock price is predicted to be increasing and when the 9 day moving average is below the 21 day moving average the stock price will be decreasing. The model used for this research is plotted very simply and more comprehensive AutoRegressive Integrated Moving Average (ARIMA) can be plotted for more comprehensively, utilising moving averages is a common methodology in technical analysis which is not the most robust for price prediction. This was created as a base comparison and for selection of key moments for analysis in the stock price movement in the machine learning model validation.

**Machine Learning Model Analysis**

A diagram of data processing

Description automatically generated

*Figure 16: Flowchart displaying the process of analysis with the machine learning model*

The Figure 16 above displays the process of the machine learning model. The compiled synthetic and real data with and without text is separately fed into the model the data is split into testing and training data which is utilised to train the model. The resulting RMSE score and the analysis conducted on the selected key moments present information that can be utilised to further the study.

**Synthetic Data**

After training the model with and without the text data the following loss charts from the RMSE data over 100 epochs have been visualised in Figure 18 in a line chart shown below. According to the loss charts the loss is reduced over time and recorded. For best practice, the model must stop training and saved prior to test loss reducing below the training loss. It can be commented that both losses decrease over time and the model learns how to predict the subsequent values. For the model without the text, the training loss for the training data started at 0.114 and finished at 0.013. The testing loss started at 0.039 and fluctuated logging the lowest value as 0.013. For the model with the text, the training loss for the training data started at 0.105 and finished at 0.012. The testing loss started at 0.035 and fluctuated logging the lowest value as 0.019. Although the model with text performed slightly better at the training data, it performed much worse in the testing data suggesting potential overfitting. The data is tabled below for observation in Figure 17.

A close-up of a data

Description automatically generated

*Figure 17: Displaying the loss result from the model training in synthetic data in a table.*

A graph of a graph of a graph

Description automatically generated with medium confidence

*Figure 18: Displaying the loss result from the model training in synthetic data.*

Even without the K fold cross validation method, according to the moving averages and visualisation of the data few key moments were selected and predicted simulating the validation method. Two key moments were selected for analysis. The first key moment displayed in Figure 19 shows the volatility highlighted with the red dotted rectangle between 2023-01 to 2023-02 for Stock 1. In the figures below the red line displays the prediction and the pink line displays how good the model performs over time.

In the synthetic data, there is a slight improvement in model with text the key moment where stock price is impacted by the very frequent tweets. It is difficult to pinpoint the exact tweet causing the volatility as multiple tweets posted on the same day is aggregated for analysis. However, looking into the Figure 20, it can be observed that during this time, tweets were mostly repeatedly negative causing the movement in the stock price.

A graph of a stock market

Description automatically generated with medium confidence

*Figure 19: Display of analysis in synthetic data, illustrating the different of the robustness of the model trained with data with text and without text.*

**A barcode with text

Description automatically generated**

*Figure 20: Displaying the tweeting behaviour of the CEO at the time of volatility in Figure 18.*

Another key moment was selected to validate if the model with the text performed better during less volatile periods. According to the Figure 21, during the period highlighted with the red dotted rectangle between 2023-9 and 2023-11 ,stock is highly correlated with the index prices. This reflects that the impact of the tweets that were synthetically generated is minimal to none. As expected, the model with the tweet performs slightly worse compared to the model without the text. It can be concluded that in periods where social media communication is less frequent and impactful the model with the text is introduced to noise reducing the robustness of the model.

To confirm the tweeting behaviour, as seen is Figure 22, the tweeting frequency is mostly positive and quite sparse in the time of the observation which validates the findings the study of the previous key moment.

A close-up of a graph

Description automatically generated

*Figure 21: Display of analysis in synthetic data, illustrating the different of the robustness of the model trained with data with text and without text.*

A bar code with green lines

Description automatically generated

*Figure 22: Displaying the tweeting behaviour of the CEO at the time in Figure 20.*

In summary in the case of synthetic data, The integration of social media analytics into these models showed improvements at certain key moments. Specifically, the models incorporating the analysis data from CEO tweets outperformed the traditional models in past case scenarios with moments that had very frequent tweeting behaviour. In addition, times where there was less tweets and volatility, model without the text input performed better than the model with the text input.

**Real Data**

In the case of the real data, similar outcomes to the loss results of the synthetic data was observed as visualised in Figure 24. In comparison, the model incorporating the real data with text performed worse in comparison to the model without text. In the case of Apple and Tim Cook it was difficult to observe significant moments for cross validation due to the lack of correlation between the price information and the tweets. However, this was expected due to the real life examples. In the case of Elon Musk, the model was not as robust as expected. A limitation can be noted, where model had to be calibrated saved and utilised more carefully. Overfitting in could also be a factor in this case. This section of the analysis requires a much more comprehensive investigation for better informed results. For the model without the text, the training loss for the training data started at 0.137 and finished at 0.015. The testing loss started at 0.070 and fluctuated logging the lowest value as 0.039. For the model with the text, the training loss for the training data started at 0.159 and finished at 0.017. The testing loss started at 0.105 and fluctuated logging the lowest value as 0.041. Although the model with text performed slightly better at the training data, it performed much worse in the testing data. Potential overfitting is observed in the model with the text. Data is displayed in a table below in the Figure 23.

A close-up of a graph

Description automatically generated

*Figure 23: Displaying the loss result from the model training in the real data in a table.*

A graph of a graph of a graph

Description automatically generated with medium confidence

*Figure 24: Displaying the loss result from the model training in the real data.*

However it can be noted that the overall performance was not robust enough to perform better at the real stock market data. In that case, additional limitations were noted, in such where the tweet data provided some noise due to its frequency, reducing the robustness of the overall prediction. The research can be furthered by filtering the tweets posted by the CEO’s prior to their incorporation into the model. As Elon Musk tweets very frequently, the additional tweet data that is not related to the company introduces noise into the model that do not have much impact on the stock price.

1. **Discussion Limitations, Challenges and Future Research Directions**

The study faced potential inaccuracies due to the influence of external factors. For example, regulatory changes, and macroeconomic indicators can impact stock prices which were incorporated by utilising three major index price data in the US. This could be further refined by conducting a corelation study. In this correlation study factors that could impact the stock prices such as interest rates, gold prices, housing prices, inflation rate could be listed and analysed to incorporate the factor that provides a result for which factor to include. Future studies could explore the impact of other social media platforms, such as Reddit and LinkedIn, on stock market trends. The development of more sophisticated and well calibrated models and real-time data processing, could further enhance predictive accuracy.

**Bibliography**

Das, A., Nandi, T., Saha, P., Das, S., Mukherjee, S., Naskar, S. K., & Saha, D. (2024, March 18). *Effect of Leaders Voice on Financial Market: An Empirical Deep learning expedition on NASDAQ, NSE, and Beyond*. arXiv.org. <https://arxiv.org/abs/2403.12161>

Fama, E. F. (1970). Efficient Capital Markets: A review of theory and Empirical work on JSTOR. *www.jstor.org*. https://www.jstor.org/stable/2325486?origin=JSTOR-pdf

Todd, A., Bowden, J., & Moshfeghi, Y. (2024). Text‐based sentiment analysis in finance: Synthesising the existing literature and exploring future directions. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *31*(1). https://doi.org/10.1002/isaf.1549

Wang, Q., Lau, R. Y. K., Xie, H., Liu, H., & Guo, X. (2024). Social Executives’ emotions and firm value: An empirical study enhanced by cognitive analytics. *Journal of Business Research*, *175*, 114575. https://doi.org/10.1016/j.jbusres.2024.114575

**Appendix**

A graph of a stock market

Description automatically generated

*Figure 1: Tesla Stock Price plotted together with S&P 500 and 9 and 21 day the moving averages.*

A graph of stock prices

Description automatically generated

*Figure 2: Apple Stock Price plotted together with S&P 500 and 9 and 21 day the moving averages.*