Sentiment Analysis to Assess the Effect of Employee Satisfaction on the Company's Stock

Inbar Shamir *

May 27, 2023

Abstract

The GlassDoor website presents an opportunity to write anonymous employee reviews about their workplace, which can be analyzed to classify them as positive or negative. Through this process, it is possible to derive an average rating of the employees' experience during a specific period of time. Based on the assumption that employee satisfaction is positively correlated with productivity, it is estimated that a positive work environment will lead to increased productivity and improved performance. Consequently, the company's products will receive more favorable reviews, potentially leading to a corresponding increase in the value of the company's stock. This paper presents a novel approach utilizing GlassDoor reviews and stocks data of nine companies to predict their future stock prices using an LSTM network. The effectiveness of the proposed approach is validated using real-world data, showcasing its performance in predicting stock prices.

1 Introduction

Stock market is a very important part of the economy. Public companies acquire capital through the stock market, and this capital funds various R&D projects to create services, products, employment which helps grow the economy. If a company performs poorly, its stock price plummets and if it performs very well, its price soars. Investors thus need to research the stock market to decide which investments will lead to profits. Predicting stock price is a very complicated task as it does not follow a fixed mathematical equation. The factors affecting the stock price of a company include financial aspects, government policies, international policies, emerging news etc.

In the last years, it was shown that employees' satisfaction is correlated with the financial state of the company, for example in the cited paper [1]. Since GlassDoor ratings are a reflection of the company culture, they impact an organization's ability to attract new talent and retain employees. Moreover, happier employees mean lower turnover, more productivity, increased creativity, and dedication, which leads to lower turnover and better company performance. When employees perform optimally, the overall company performance increases, leading to higher productivity that ultimately impacts the bottom line.

Sentiment Analysis can help us figure out the mood and emotions of general public and gather insightful information regarding the context. Sentiment Analysis is a process of analyzing data and classifying it based on the need of the research. These sentiments can be used for a better understanding of various events and impact caused by it. In this paper, sentiment analysis was applied to GlassDoor reviews using the TextBlob library, to derive a sentiment score. This score was then utilized to enhance the prediction of stock closing prices.

I used an LSTM (Long Short-Term Memory) model to forecast stock prices by incorporating both stock data and reviews scores. LSTM is a type of recurrent neural network (RNN) that excels in capturing long-term dependencies in sequential data. It utilizes a memory cell and specialized gates to selectively retain or forget information over time, enabling it to effectively learn patterns and make accurate predictions.

This paper is structured as follows: in section 2, I provide an overview of the related literature on sentiment analysis, stock prediction, and the use of LSTM models. Section 3 describes the methodology employed, including the data collection process, sentiment analysis techniques, pre processing and the

*Dept. of Computer Science, Holon Institute Of Technology, Holon, Israel

Company	Number of Reviews	Stock Name
IBM	20,329	IBM
Oracle	11,541	ORCL
McDonald's	9,212	MCD
Microsoft	7,441	MSFT
JP Morgan	7,348	JPM
SAP	6,072	SAP
Apple	6,053	AAPL
Citi	5,456	C
Google	4,182	GOOG

Table 1: companies in final data set

implementation of the LSTM model. The experimental results and analysis are presented in section 4, followed by a discussion of the findings in section 5.

2 Related Work

There has been a number of efforts to predict stock prices based on textual information. Researchers have employed NLP (natural language processing) and machine or deep learning techniques to analyze sources like news articles, social media data, and corporate filings. The aim is to improve prediction accuracy by identifying patterns and factors that influence market trends. The author of this article [2] for example, performed sentiment analysis on news articles and created an LSTM-RNN model to predict the direction of the stock movement. Akita et al. [3] proposed a similar approach to predict stock prices by employing distributed representations of news articles and considering the correlations between multiple companies within the same industry. This paper [4] used the pre-training language representation model BERT to perform sentiment analysis on collected news articles and PTT forum posts related to individual stocks. They calculated the daily sentiment probabilities together with historical stock transaction data as input vector to LSTM neural network to forecast next day stocks opening price. Halder [5] explored different Deep Learning techniques to predict stock price. He proposed a successful FinBERT-LSTM model. This model incorporates news sentiments into an LSTM network to make predictions. FinBERT is a pre-trained BERT model fine-tuned for financial sentiment.

Several studies have demonstrated a correlation between employee satisfaction and stock performance. Specifically, a 2017 study by Symitsi et al. [1] used more than 326,000 GlassDoor reviews to quantify the impact of employee satisfaction on the long-run stock performance of companies. According to the study, "employees' online reviews are good predictors of a firm's financial results".

Furthermore, considerable research has been conducted on sentiment analysis of GlassDoor reviews. For instance, a notable study [6] aims to extract significant aspects from user reviews and classifying the overall sentiment associated with them.

This study shows a novel approach of combining financial historical information with sentiment analysis scores obtained from GlassDoor reviews. By integrating these two sources of data, the study aims to achieve improved results in predicting stock prices. This innovative approach harnesses the power of both numerical data and textual sentiment analysis to enhance the accuracy and effectiveness of stock price forecasting.

3 Methods

3.1 Data Sets

For the job reviews from GlassDoor website, I downloaded the following data set from Kaggle: GlassDoor Job Reviews. I took the columns: $Firm, Date\_Review, Current, Pros, Cons$ of the nine companies with the largest number of current employees' reviews from years 2015 to 2019, that has finance data from Yahoo Finance - see Table 1. The columns I use for the finance data are: $Date, Open, High, Low, Close, AdjClose, Volume$.

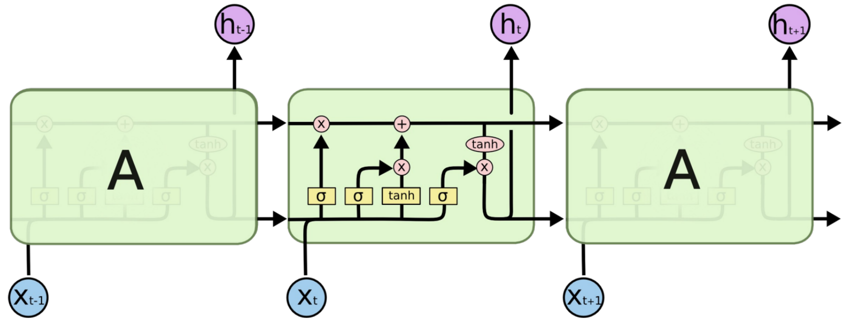

Figure 1: LSTM structure

3.2 Sentiment Analysis using TextBlob

TextBlob is a python library for Natural Language Processing (NLP). For sentiment analysis, TextBlob returns polarity and subjectivity of a sentence. For the reviews' analysis I referred only to the polarity. Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment.
The reviewers in GlassDoor website can write the advantages and disadvantages of working in the company according to their feelings. I used TextBlob to extract the polarity of the pros part of the review, and for the cons part. In order to get a positive number between 0 to 1, I normalized the polarity as follows:

$$score = \frac{polarity + 1}{2}$$

For the final analysis, I calculated the average of both polarities score, and saved it to the relevant row in the data set. A final score of 1 means that both pros and cons were analyzed as positive, so the overall feeling of the reviewer is positive. As the score approaches 0, it means that both pros and cons were analyzed as negative, therefore the reviewer probably has a negative feeling about working in the company.

3.3 Data Aggregation

I combined the stock data and sentiment analysis scores for each company based on the dates, resulting in 9 data sets containing both stock data and sentiment scores. In case of more than one review of the same company in one day, I grouped the scores by taking their mean value and save it as the score of this date. Due to the presence of daily stock data but not daily reviews, there are some missing values in the merged data set (less than 5% for each data set). To address this, I utilized the fillna() function to fill the missing values with the mean values of the scores from seven days before. This approach was chosen under the assumption that employee sentiments tend to remain relatively constant over a few consecutive days. In summary, the parameters of the model are: open price, close price, high price, low price, volume and sentiment score according to the dates. The target vector is the close price of the stock for each day in the range.

3.4 Pre Processing

Firstly, I divided the data into two sets considering the chronological order of the data: the train set contained all the data from 2015 to 2018, and the test set included only the data of 2019. As a result, the train-test ratio equals 0.25. I used the close price column as the target variable vector, so the model needs to predict its value in the next day using its former values. Then, I performed data pre-processing, which involved normalizing the numerical features (Open, High, Low, Close, Adj Close, Volume, Sentiment Score) using Min-Max scaling. This process ensured that all the features were brought to a similar scale, where 0 represented the lowest price/score and 1 represented the highest price/score. Finally, after scaling and reordering the data to ensure an equal number of rows in the train and test sets for each company, I organized the train and test sets to align with a window of 30 days (30 features) to the neural network. This configuration allowed the LSTM model to learn from the 30-day window, enhancing its ability to predict patterns within the data set.

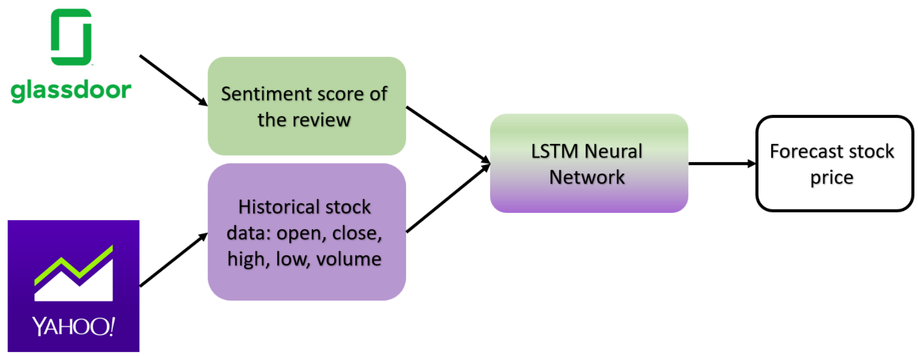

Figure 2: Model Architecture

3.5 LSTM Model

LSTM or Long Short-Term Memory, is a type of gated Recurrent Neural Network (RNN), that initially proposed in 1997. Unlike traditional RNNs that struggle to maintain information over long time gaps, LSTMs are designed to overcome this limitation. By incorporating specialized memory cells and gating mechanisms 1, LSTMs can capture long-term dependencies in sequential data and make more accurate predictions, making them better suited for predicting stock prices. The LSTM network I constructed is composed of 4 layers: 1 input, 2 hidden, 1 output. The input layer accepts input of the shape (30, 7). For the hidden layers I used LSTM layers with a "tanh" activation function. The input and hidden layers consist 50 neurons each, and the output layer is Dense with one neuron. I added 3 dropout layers in between them, with dropout rates 0.08, 0.05, 0.01 respectively, to avoid over-fitting. I trained the LSTM network with the nine train sets I prepared, for 6 epochs with a batch size of 55, using MSE loss function and Adam optimizer having its default 0.001 learning rate. Overall, The model's architecture can be visualized as follows 2.

4 Experimental Results

I evaluated the model respectively using the nine test sets, and calculated the MAE and MAPE for each company:

- Mean Absolute Error – Mean Absolute Error (MAE) is the mean of the distance between the target variable and predicted value. MAE can be computed by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Absolute Percentage Error – Mean Absolute Percentage Error (MAPE) is the mean of the distance between the target variable and the predicted value as a fraction of the target variable. MAPE can be computed by:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

I repeated the entire process twice, once incorporating the sentiment analysis scores and once without it. The results for both scenarios are recorded in the table 2. Observing the table, it is evident that including the sentiment analysis scores on GlassDoor reviews led to an improvement in the results. Specifically, the average MAPE without the sentiment analysis scores is approximately 0.0028 larger than the MAPE when considering them. Out of all the nine companies examined, IBM was the only one for which the results did not show improvement when incorporating the reviews data. The most favorable outcomes among the companies analyzed were exhibited by Google. This is evident in the graph 3 describing the relationship between the actual close prices and the predicted close prices of Google stock over time. By observing the second graph compared to the first, we can see the superior

Stock	MAE	MAPE	MAE (no sentiment)	MAPE (no sentiment)
AAPL	1.142615	0.020240	1.293503	0.022648
C	1.165517	0.017368	1.178460	0.017520
GOOG	0.854721	0.014324	0.963779	0.016038
IBM	4.058309	0.030949	2.505354	0.019144
JPM	2.034864	0.017243	2.420242	0.020244
MCD	3.307653	0.016078	5.034764	0.024475
MSFT	4.400934	0.031577	6.485611	0.046658
ORCL	0.801892	0.014538	1.005687	0.018179
SAP	2.003409	0.016013	2.382367	0.018751
Average	2.196657	0.019814	2.585530	0.022629

Table 2: final results

performance of the model when sentiment analysis scores are incorporated compared to when they are not included. However, in comparison to the article by Halder [5], which employed a similar LSTM model using news articles and achieved a MAPE score of 0.014568, my average MAPE score is higher and equals 0.019814.

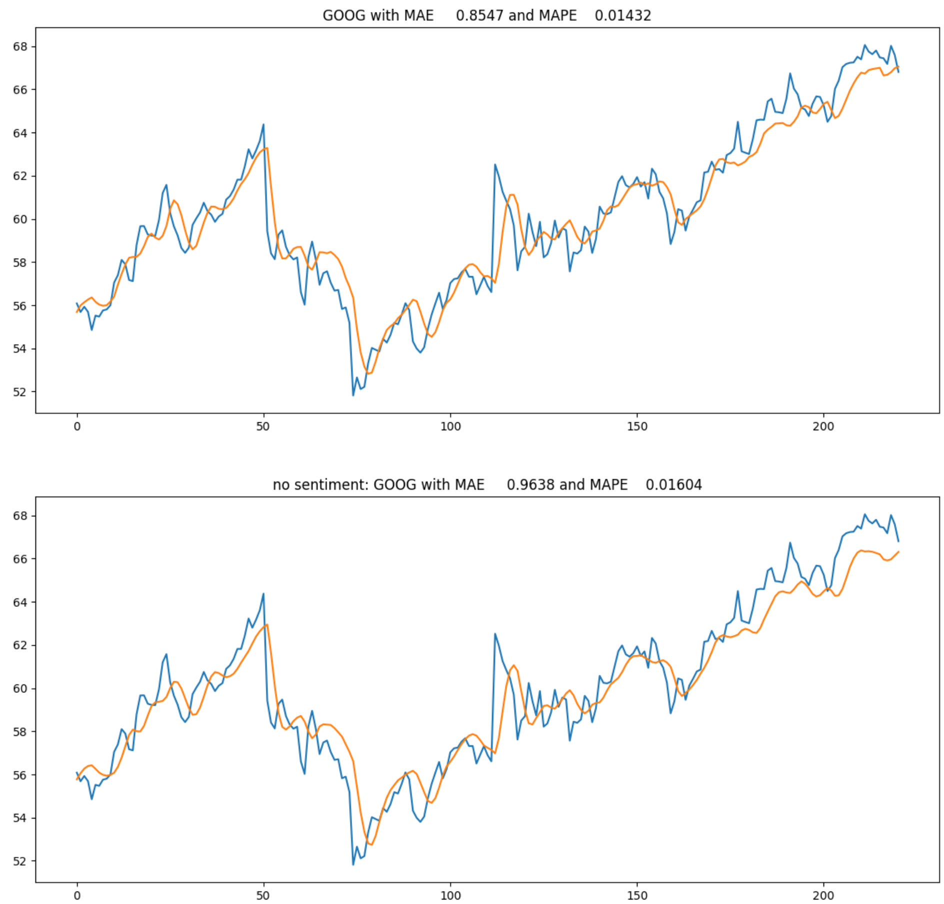

Figure 3: Google stock predicted & true close prices over time, including/excluding sentiment scores

5 Discussion

Predicting stock price is a very challenging task. In this study, the main focus was to explore the potential contribution of employees' reviews in improving stock forecasting for nine companies. To achieve this, I performed a sentiment analysis of reviews from GlassDoor website by using TextBlob library. I built an LSTM neural network that was then trained using both the historical stock data obtained from Yahoo Finance and the sentiment analysis scores. The goal was to leverage the valuable insights from employee reviews to enhance the accuracy of stock price predictions. The results demonstrated the significance of incorporating sentiment analysis in the forecasting process, with an average MAPE score of 0.019814, in contrast to average MAPE of 0.022629 of a baseline method. While this performance is not as low as the MAPE score of 0.014568 achieved by Halder et al. [5] using a similar LSTM model with news articles, it still showcases the potential value of employee reviews in stock forecasting. These findings highlight the importance of further research and optimization to refine the approach and explore alternative techniques for even better prediction results.

References

[1] Efthymia Symitsi, Panagiotis Stamolampros, and George Daskalakis. Employees' online reviews and equity prices. *Economics Letters*, 162:53–55, 2018.

[2] D. Kumar. Stock forecasting using natural language and recurrent network. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, 2020.

[3] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara. Deep learning for stock prediction using numerical and textual information. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016.

[4] C. R. Ko and H. T. Chang. Lstm-based sentiment analysis for stock price forecast. *PeerJ. Computer science*, 7:e408, 2021.

[5] Shayan Halder. Finbert-lstm: Deep learning based stock price prediction using news sentiment analysis. *arXiv preprint arXiv:2211.07392*, 2022.

[6] Hrithika Yadav, Kartik Dwivedi, and G. Abirami. Sentiment analysis of company reviews using machine learning. In *2022 3rd International Conference for Emerging Technology (INCET)*, pages 1–5, 2022.