An Innovative Artificial Intelligence and Natural Language Processing Framework for Asset Price Forecasting Based on Islamic Finance: A Case Study of the Saudi Stock Market

Klemens Katterbauer[*♣] and **Philippe Moschetta**[♣]

[♣]Euclid University, Central African Republic

Submitted: May 29, 2021 • Accepted: December 5, 2021

ABSTRACT: Artificial intelligence has transformed the forecasting of stock prices and the evaluation of companies. Novel techniques, allowing the real-time processing of large amounts of data, have enabled the use of data on various external factors to improve the forecasting of the company's value and stock price. Although conventional approaches solely focus on the use of quantitative data, history has shown that news announcements and statements may significantly affect the performance of the stock value of companies. We present an innovative framework for integrating a nonlinear autoregressive network with a natural language processing approach to analyze stock price movements and forecast stock prices. The framework analyzes and processes the company's financial statements, determining indicative factors and transforming them into categorical parameters which are then integrated into a nonlinear autoregressive network to estimate and forecast the company's stock price. The analysis of several Saudi companies listed in the Tadawul index affirms the improved estimation of the stock price and the possibility of a more precise prediction of long-term stock price evolution.

JEL classification: G17, C70.

Keywords: Artificial Intelligence, Islamic Finance, Financial Statements, Natural Language Processing, Saudi Arabia, Financial Markets.

[*]Corresponding Author. E-mail: katterbauer.klemens@gmail.com

DOI: 10.2478/erfin-2021-0009

1 Introduction

Forecasting of both individual company's stock price performance and that of the broader market has been significantly challenging due to the multitude of factors that can influence the overall performance of stock prices. This issue has encouraged the development of multiple models and approaches that differ in terms of the time horizons considered and whether a qualitative or quantitative approach is used. We start with a comprehensive literature review related to stock price performance forecasting and its impact on Islamic finance.

1.1 Literature Review

Quantitative approaches to stock price forecasting can be divided into stochastic, statistic, and deterministic models. Stochastic models have been widely developed to address the randomness in the nature of stock price movements, which may follow random probability distribution patterns similar to erratic behaviors such as the Brownian motion. These models typically encompass a stochastic process optimized via estimation of the stochastic parameters, such as the mean, the drift, and the volatility of the process (Farida et al., 2018). Geometric Brownian motion models have been attractive to use, given the relative ease of parameter estimation and computational efficiency. Nevertheless, these models face considerable challenge due to the fact that they may work in the short term but fail to capture trends or deviations from a purely stochastic behavior of the stock.

Statistical approaches such as time-series analysis with ARMA models and their variations have been widely used to capture data trends and perform economic forecasting over multiple years (Chi, 2018). These models have proven to be relatively efficient and robust for short-term price movement modeling (minutes to a few days), however, long-term prediction (several days to months) has been a challenge due to the multiple additional factors that may affect the forecasts and which may not be easily captured in these models. Specifically, statements made by individuals and expert interpretations are challenging to incorporate into these models without a semantic analysis model.

Novel artificial intelligence techniques have had a significant impact on improving the forecasting of stock prices in recent years and have been extensively applied to a variety of different problems. Vijh et al. (2020) used a random forest approach for a prediction of stock closing prices for 10 major stocks. The utmost challenge that arises from modeling is that high and low closing prices for each day are available, in addition to moving averages for 7, 14, and 21 days. Given the bounds for each day, artificial neural networks performed best for the estimation of the closing prices. However, the results indicate a rather limited ability to extend the technique to stock price forecasting over the long term, as these estimates are solely conducted on a day-to-day basis.

Another machine learning approach was introduced by Guo et al. (2018), who developed an adaptive support vector machine regression model to satisfy the changing data flow for high-frequency stock price forecasting (Guo et al., 2018; Wang and Wang, 2015). The timescales for the model were set as daily, 30-minute, and 5-minute. The dynamic optimization based on the different timescales and using particle swarm optimization exhibited stronger performance.

Although data-driven models may perform well for short-term stock price forecasting, long-term stock price forecasting represents a completely different challenge given the abundance of other factors that must be considered.

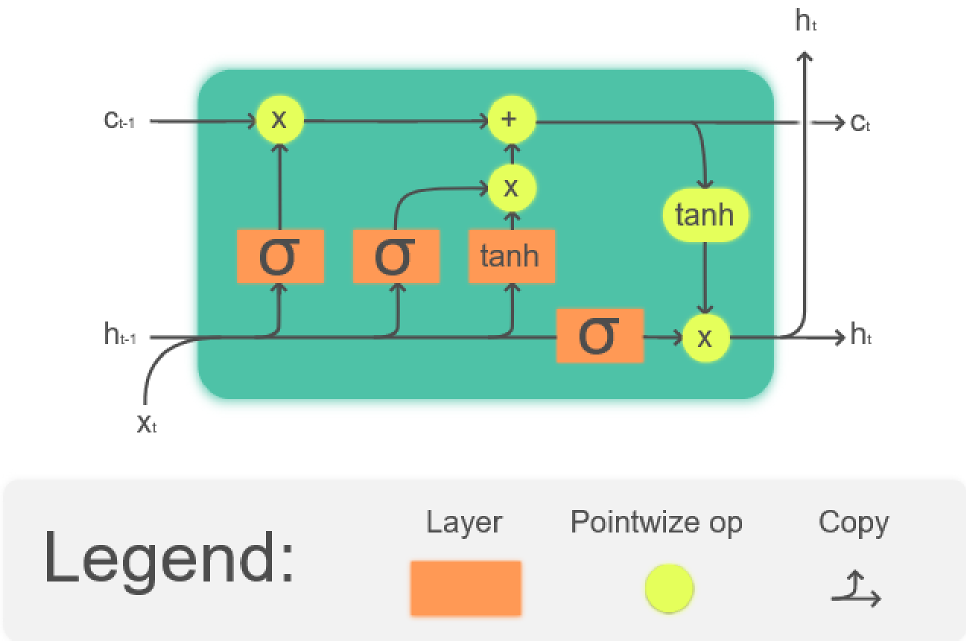
Recent advances in natural language processing (NLP) have allowed for efficient and effective analysis of textual data to infer both content and intent. NLP is concerned with the interaction of computers and human language, allowing computers to understand and interpret, as well as respond to, human language. Although symbolic NLP has been known since the 1950s, the advances in statistical NLP that began in the early 1990s have brought about the breakthrough of neural NLP. Symbolic NLP had a significant drawback because it was dependent on specific sentence and word patterns, which created challenges due to ambiguities in language and the various forms in which humans could voice their opinions and views. Neural NLP has significantly expanded the usability of machine learning for determining the intent of human language and has adapted more easily to various human language patterns.

Stock price forecasting using NLP has only recently become of interest. Mehtab and Sen (2019) developed the long-short term model (LSTM) for predicting the closing price of a stock with a time horizon of one week. Furthermore, the authors incorporated Twitter sentiment analysis data in the forecast (Mehtab and Sen, 2019; Zhang et al., 2019). The approach of taking into account views and thoughts on Twitter is interesting, but the correlation between textual information on Twitter and stock price movements is very limited. Moreover, Twitter sentiment may be rather short-term and potentially incomplete. As outlined in Figure 1, LSTM uses several sigma activation functions as well as a *tanh* function to combine the input vector X_t with the cell state vector c_t and hidden state vector h_t in order to determine the cell state and hidden state vector at the next time step. LSTM units solve the vanishing gradient problem that is commonly encountered in recurrent neural networks (RNN) by allowing the gradients to flow unchanged. This makes the LSTMs attractive in time series problems as they are generally more robust than their RNN counterparts.

Financial and annual reports have long presented an important insight into the financial conditions of a company and its long-term outlook. However, although extensive research on human interpretation of financial reports has been conducted within the last decades, limited work on the automatic interpretation of the reports has been performed.

Lewis and Young (2019) present an extensive overview of the opportunities for implementing NLP in the analysis of financial accounting and reporting statements. Whereas balance sheets and cash flow statements may provide metric information on the status of a company, textual information within the financial reports has exhibited crucial insights

Figure 1: Graphical illustration of a typical LSTM framework



into the performance of the company. An important challenge in analyzing financial reports is a significant number of specialized words and complex sentence structure. As compared to conventional text language, such as on Twitter, financial reports use very concise language and sentence corrections or incorrectly written words are seldom a challenge. However, understanding the intent and connecting the various contexts together is more challenging. Social media based text interpretation may emphasize sentiment analysis, with only a few classification options, whereas analysis of financial reports requires multi-class classification and refined multi-page understanding (Lewis and Young, 2019).

1.2 Research Outline

We present a novel framework for stock price forecasting using an artificial intelligence framework for the integration of stock price and textual information from financial reports. The framework distinguishes itself from existing frameworks by integrating textual information from financial reports with a specific focus on Islamic finance principles. We tested the framework on data from companies on the Saudi stock exchange and were able to demonstrate strong estimation and forecasting performance. The framework provides efficient stock price forecasting and incorporates information related to Islamic finance principles that may provide better guidance for stock price performance on Islamic stock exchanges such as the Tadawul stock exchange. Furthermore, it can be used to determine the effect of the text components related to Islamic finance on the estimation of performance.

2 Methodology

We developed an innovative LSTM-NLP framework that integrates both historical stock price information and data extracted from financial reports to forecast the stock price of a corporation. The NLP component incorporates a PDF text reader that is connected to a text pre-processor. The text pre-processor removes any pages and textual components that contain solely legal and administrative information, such as the company contact information, corporate governance regulations, and board member information, and are relatively unimportant in determining the company's future performance. The processed pages are then separated into interpretative textual information and numeric data. The numeric data by itself incorporates a significant amount of information; however, the NLP in this approach is used only on the textual information, which includes an interpretation of the numeric data. This is done to avoid any potential conflict between the numerical data and the textual interpretation of that data. The remaining textual information is then tokenized, and the individual sentences are separated. After the sentence breakdown, part-of-speech tagging and chunking are performed. In doing so, the main textual components are quantified in terms of what actions were taken or what situation the company encountered. This is followed by chaining, where related sentences are connected. The chaining plays a crucial role in ensuring that multi-sentence intents are adequately considered.

The resulting components are then compared with positive and negative statements related to stock price movements. For example, a statement such as "We expect an increase in revenues and profits over the coming years" is scored as a more similar to a positive statement on a stock development, in contrast to a statement such as "There will be considerable risks in the coming years". For each report, the ten highest scores for each of the indicative statements are then averaged and incorporated into the LSTM framework. The date assigned to the statement in the framework is the date of publication of the financial report for the specific year, and the score is maintained for the rest of the year. The data are integrated into the LSTM framework as differential values - in other words, the value is set to the similarity score on the date when the financial report was published and then equal to zero afterwards. The assumption is that the LSTM experiences a memory loss, which implies that the effect is felt solely at a specific point in time.

LSTM networks have become an attractive alternative to machine learning models for sequences of data such as videos and time series. In contrast to standard feedforward neural networks, LSTM networks incorporate feedback connections that capture memory of the behavior of the data. The LSTM is composed of a cell, input gate, output gate, and forget gate. The cell plays a crucial role in that it remembers the values of a specified time interval, while the gates regulate the flow in and out of the cell.

Although LSTM networks are recurrent neural networks, they overcome the vanishing gradient problem since the gradients remain to flow unchanged. Specifically, during the error-value backpropagation, errors are kept within the cell. This allows for a continuous feeding back of the errors to the gate until there is sufficient learning that these values can be cut off. However, there may still exist challenges, related to the exploding gradient problem, that need to be addressed.

The LSTM has proved to be a promising option for dealing with time series forecasting that is able to account for the jump condition for the input values arising from report data integration.

3 Results

We evaluated the performance of the framework on ten major stock exchange listed corporations in Saudi Arabia, listed in Table 1 (Saudi Exchange, 2021). The companies were selected according to three criteria. The first criterion was that the stock had sufficient daily trading volume. This is important to avoid having liquidity effects leading to erratic stock price behavior. The average daily traded volume is a good indicator of the liquidity of the stock in the market and can be used to assess whether there are any challenges with respect to the purchase and sale of the stock. The second criterion was that there should have been an adequate distribution of the stocks across various sectors. Banking and energy are the main sectors represented on the Tadawul stock exchange, given their importance for the Saudi economy, and the aim was to have companies from multiple sectors for the assessment of stock price forecasting. The third criterion was that all of the companies should have had financial reports for the previous four years published. This was required so that we had the same number of financial reports available for each company included in the evaluation.

Table 1: Companies used in the evaluation of the LSTM-NLP framework

Corporation	Sector	Average Daily Traded Value (in million SAR)
Alinma Bank	Financials	46
Almarai Co.	Consumer Staples	10
Al Rajhi Bank	Financials	112
Bupa Arabia for Cooperative Insurance Co.	Financials	2
Etihad Etisalat	Communication Services	41
Jarir Marketing Co.	Consumer Discretionary	23
Najran Cement Co.	Materials	3
National Petrochemical Co.	Materials	28
Rabigh Refining and Petrochemical Co.	Energy	21
Saudi Arabian Mining Co.	Materials	13

Note: Companies were selected from the Tadawul stock exchange based on their market importance and to achieve diversity across sectors.

The stock price was analyzed from January 2016 until April 2021. High and low closing stock price were taken in 7-day intervals. For each time step, we forecasted the stock price seven days ahead, where besides the previous stock price, the financial report information, and the publication intervals of the report, was incorporated. Given the difference in data integration intervals, the weighting of the input parameters for the NLP-derived data needed to be optimized to avoid overfitting. Namely, given the relative sparsity in time for these data, NLP-derived data needed to be weighted more strongly in comparison to stock price data.

Figure 2 presents the comparison of the predicted stock prices versus the actual stock prices of all the analyzed companies.

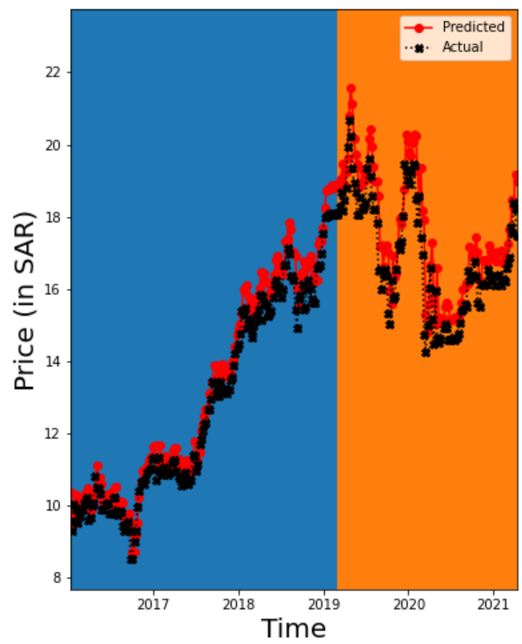
Strong training and forecasting performance, demonstrated by low root mean square error (RMSE) value (below 2%) for all stock prices, can be seen. The training period covered the time from 2016 until 2019, while the forecasting period was extended until April 2021. As it can be observed, the stock price predictions in both training and forecasting periods are highly accurate, capturing the main dynamics of the stock price evolution. Even out-of-distribution and outlier stock prices are well captured. For machine learning and artificial intelligence applications, capturing outliers accurately is a major challenge in time series forecasting. Conventional time series modeling frameworks, such as ARIMA models or nonlinear regression methods, have limited ability to track strong nonlinearities in the data. This results from the limitations in complexity and adaptability of these models which possess only a certain number of parameters that can be fine-tuned. Considering that outliers are extreme events that occur with a small probability, they are typically captured only to a limited extent, given the reduced weight in the loss function. LSTM-based deep learning models are well suited to capture these outliers accurately in the time series data.

The previous results outlined the predictive performance of the framework; however, they provide relatively little insight into which components of the input data have the most significant effect on the framework performance. In figure 3, we provide the force plots of the SHAP (SHAPley Additive exPlanations) values for the various stock prices. SHAP provides a framework for determining why a prediction of the model is different from a baseline. The total difference is subdivided into the SHAP values for the individual features. The feature values that increase the accuracy of the prediction are presented in pink and those that decrease it are indicated in blue. The difference between the lengths of the blue and pink bars is the prediction value. SHAP values provide a rather robust framework for determining the effect of various features on the prediction performance of the LSTM stock price estimation. A closer analysis of the results indicates that the various input factors such as the stock price and the NLP-derived similarity scores have different impact factors for the various stock price estimates. Each similarity score is a numeric representation of a verbal (or textual) indication of the anticipated development of the stock price, such as “the stock price is increasing” or “the stock price is stable”.

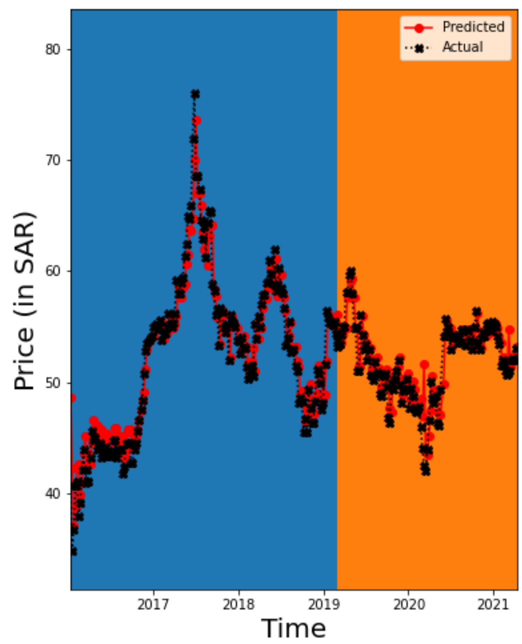
The fact that the values represent solely the effect of the specific feature in question is crucial for the interpretation of the SHAP values. In the case of the stock Alinma, both SimScores 1 and SimScores 3 represent statements of the expectation of a higher future

Figure 2: Predicted versus actual stock prices for each company

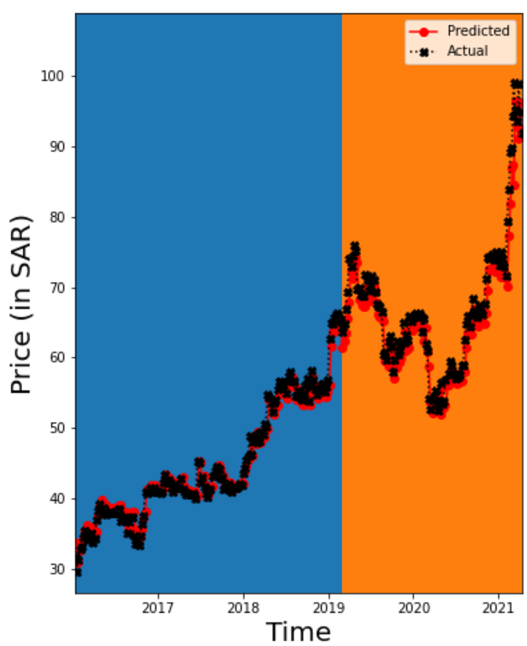
(a) Alinma



(b) Almarai



(c) Al Rajhi



(d) Bupa Arabia

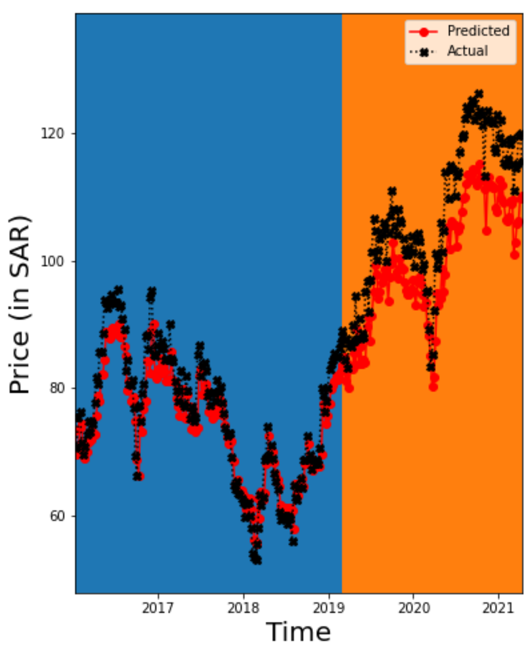
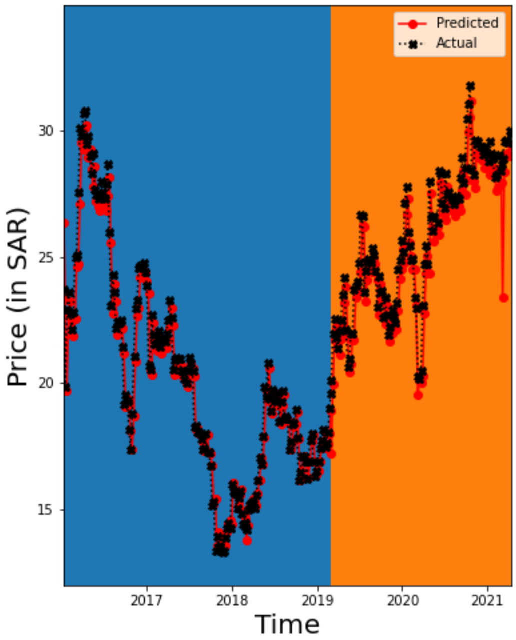
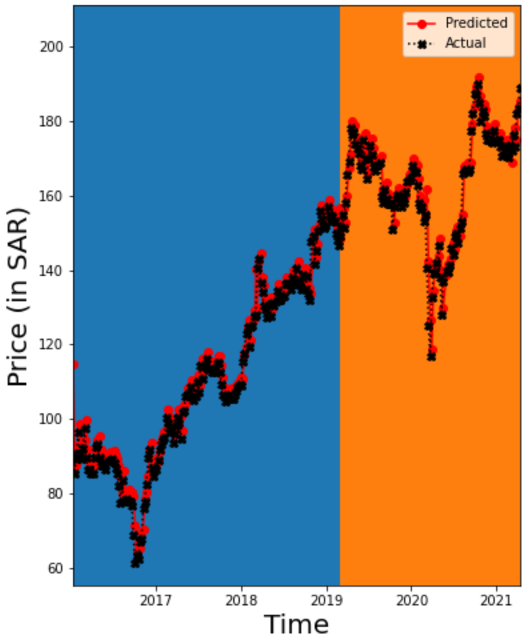


Figure 2: Predicted versus actual stock prices for each company (Continued)

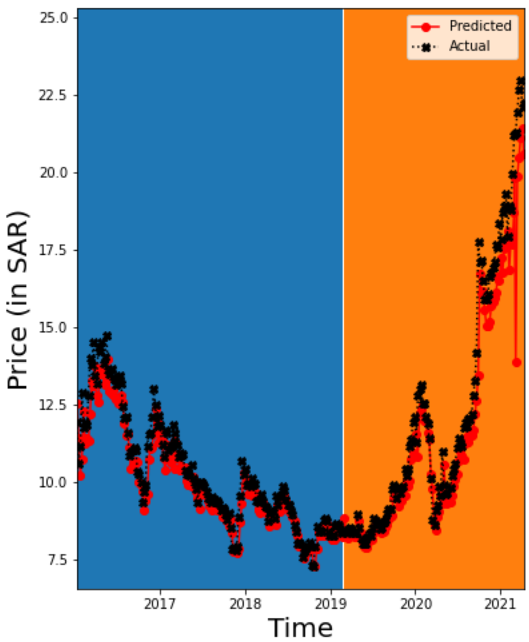
(e) Etihad Etisalat



(f) Jarir



(g) Najran Cement



(h) Petrochem

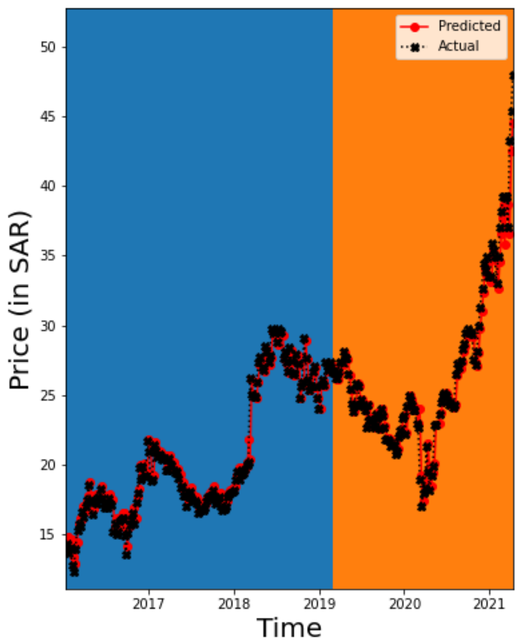
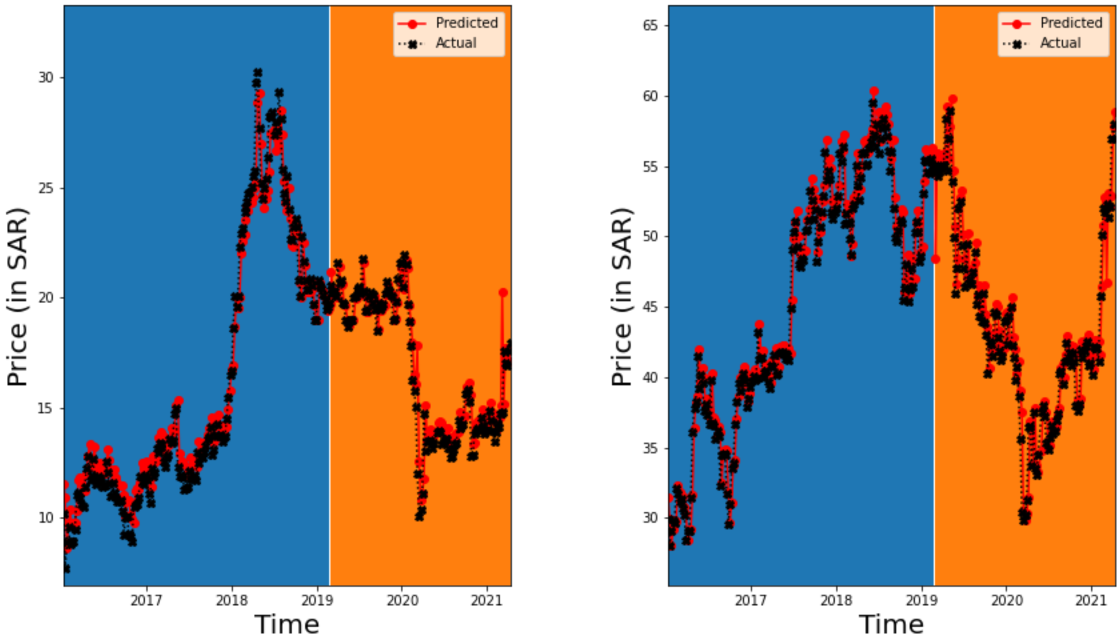


Figure 2: Predicted versus actual stock prices for each company (Continued)

(i) Petro Rabigh (j) Saudi Mining



Note: The training period is indicated in blue, and the test period in orange.

stock price, which is observed in the general increasing stock price over the observation period. Likewise, SimScore 4 represents a statement of a trend towards a lower stock price. The effect can be observed from the fact that the stock price went shortly lower, before recovering again. For Najran Cement, previous stock prices alone affected the forecast of its stock price, and the NLP data had only a minor effect on its forecast. When considering the overall stock price for the company, one cannot observe a particular characteristic for this occurrence. However, a closer look at the similarity scores for the stock price development and the sharp stock price increase in the last several years, reveals that the stock price was overweighted. This is also observable in the financial report that contains both negative and positive stock price outlooks, with a rather neutral behavior. The inputs into the model led to a prediction of stock prices staying relatively level-stable, but the actual prices increased.

An important element of the framework is the use of NLP to take into account references to Islamic finance options, such as Murabaha deposits and other products connected to it. This type of information is relatively rare in these financial reports; however, when they occur, they provide a good indication of additional revenue streams the corporation has available to boost its earnings and increase its stock price.

4 Conclusion

Artificial intelligence has played a prominent role in the last several years in enhancing the financial modeling and forecasting of financial instruments. Deep learning frameworks have the advantage that they can accurately model complex system behaviors, making them well-suited for stock price forecasting. Although numerical forecasting has been extensively researched, the incorporation of textual components, such as from the company's financial reports, is a more recent addition. Textual information from company reports and social media has considerable value for determining market sentiment and general outlook for stock prices. We have presented an innovative LSTM-NLP framework for the forecasting of stock prices in Saudi Arabia, including the ability to consider Islamic finance concepts. The framework was examined on ten companies listed on a Saudi Arabian stock exchange and found to exhibit strong forecasting performance for the stocks. A feature importance analysis demonstrated the significant effect of textual information on enhancing the forecasting of stock prices. Future research may use NLP to investigate financial statement information related to specific Islamic forms of liabilities and the outlook of the corporation. A detailed analysis of this type of textual information may allow automatic report analysis and interpretation within the context of Islamic finance and enhance AI-driven investment decisions.

References

Chi, W. L. (2018). Stock Price Forecasting Based on Time Series Analysis. *AIP Conference Proceedings*, 1967(1). https://doi.org/10.1063/1.5039106.

Farida, A., Affianti, I., and Putri, E. (2018). Stock Price Prediction Using Geometric Brownian Motion. *Journal of Physics: Conference Series*, 974.

Guo, Y., Han, S., Shen, C., Li, Y., Yin, X., and Bai, Y. (2018). An Adaptive SVR for High-frequency Stock Price Forecasting. *IEEE Access*, 6:11397–11404.

Lewis, C. and Young, S. (2019). Fad or Future? Automated Analysis of Financial Text and Its Implications for Corporate Reporting. *Accounting and Business Research*, 49(5):587–615.

Mehtab, S. and Sen, J. (2019). A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing. Available at SSRN: https://ssrn.com/abstract=3502624.

Saudi Exchange (2021). Market Participants - Market List. Retrieved from Saudi Exchange: https://www.saudiexchange.sa/.

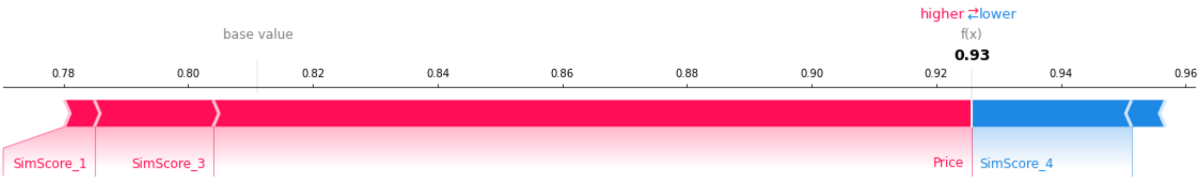
Vijh, M., Chandola, D., Tikkiwal, V. A., and Kumar, A. (2020). Stock Closing Price Prediction Using Machine Learning Techniques. *Procedia Computer Science*, 167:599–606.

Wang, J. and Wang, J. (2015). Forecasting Stock Market Indexes Using Principle Component Analysis and Stochastic Time Effective Neural Networks. *Neurocomputing*, 156:68–78.

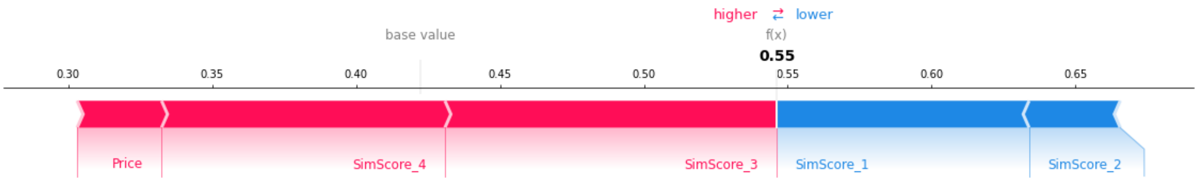
Zhang, J., Teng, Y.-F., and Chen, W. (2019). Support Vector Regression With Modified Firefly Algorithm for Stock Price Forecasting. *Applied Intelligence*, 49(5):1658–1674.

Figure 3: Effect on the prediction performance of the LSTM-NLP framework

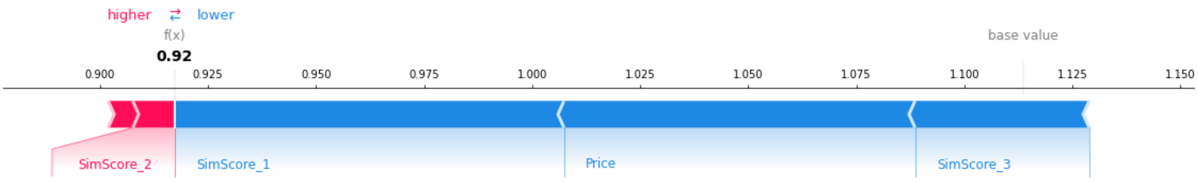
(a) Alinma



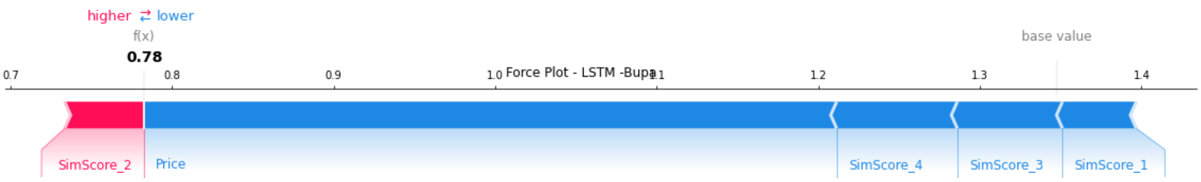
(b) Almarai



(c) Al Rajhi



(d) Bupa



(e) Etihad Etisalat

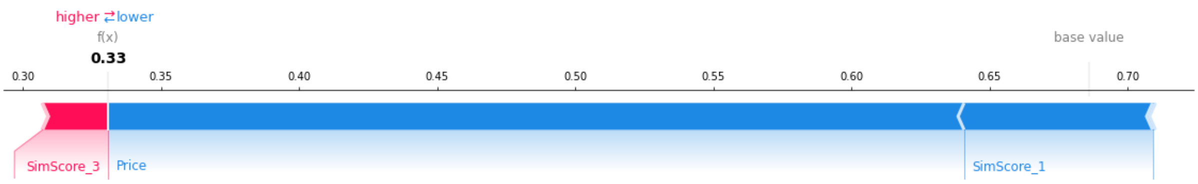
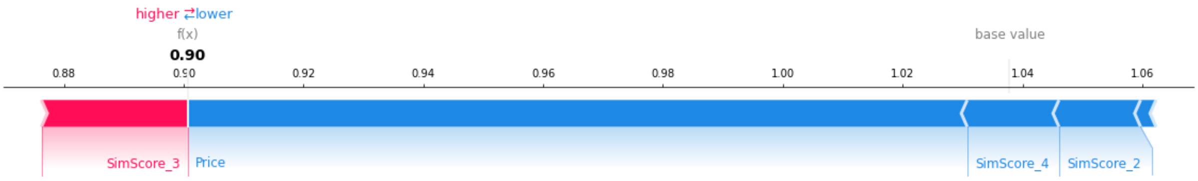
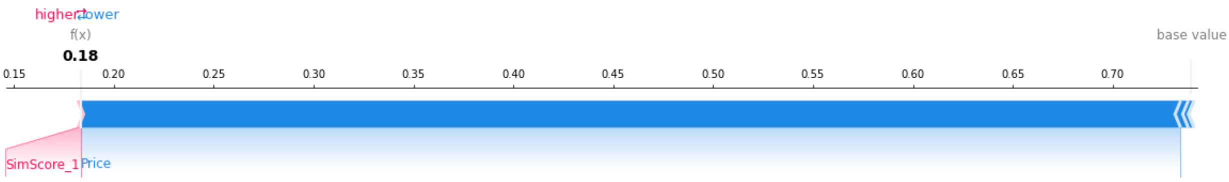


Figure 3: Effect on the prediction performance of the LSTM-NLP framework (Continued)

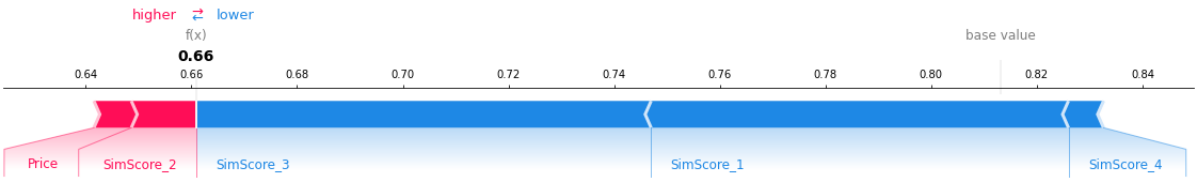
(f) Jarir



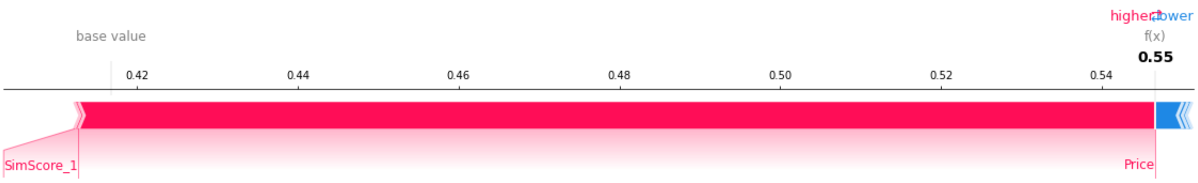
(g) Najran Cement



(h) Petrochem



(i) Petro Rabigh



(j) Saudi Mining