EconoMind: Navigating Your Financial Future

Mohammed Alkaldi
KAUST Academy
Mohammed.kaldi@kaust.edu.sa

Talal AlKharashi
KAUST Academy
talal.kharashi@kaust.edu.sa

Turki AlOtaibi
KAUST Academy
turki.otaibi@kaust.edu.sa

Nawaf AlDowayan
KAUST Academy
nawaf.dowayan@kaust.edu.sa

August 2023

Abstract

The understanding of the stock market field is very essential to Saudi 2030 vision, and in efforts to accomplish that, one of the KPIs to ensuring education about that field is by optimizing the number of responsible investors in the market. However, there has been too many stigmas for the public investors that made them deflect to the real-estate business; low-risk, low-reward; over a period of 30-40 years; which yields 1-3 percent annually on average. Providing the means to educate individuals about the stock market and financial information is very crucial. In our approach we used a multi-model architecture to encapsulate textual news and numerical financial information to better predict the stock price. Doing so, will identify the leading variables towards defining what makes the price behave in the way it does. With an LSTM-based model for time-series data, and a Transformer-based model for textual data, the performance was adequate (68 percent accuracy for the sentiment analysis model and 1.038 Riyal error in the time-series model). Further exploration of feasibility of using Transformer-based architecture for time-series is recommended.

1 Introduction

One of the pillars of Vision 2030 is to have a knowledgeable and financially responsible community [1]. Many individuals shy away from the stock market due to a combination of factors. A significant barrier is the perceived complexity and lack of understanding of how markets operate, making potential investors wary of making uninformed decisions. This is further exacerbated by memories of past market crashes, which instill a fear of potential financial loss. Additionally, mistrust of financial institutions, stemming from past scandals and unethical behaviors, makes some hesitant to entrust their money to these entities. Cultural beliefs might equate stock investing to gambling, and for others, immediate financial obligations or a preference for tangible assets like real estate take precedence. Lastly, the sheer volume of investment choices can be overwhelming, leading to decision paralysis, while in some regions, the lack of a robust financial infrastructure discourages participation.

Stock market investment has outperformed any other form of investment over the past two centuries [2]. Current forms of investments done in the Saudi community are classical and to some "old school"; buying an estate land and re-sell it after 30-or-so years, which in the best case scenario yields a growth of 6 percent annually. Conversely, the same invest-

ment if done on the Saudi stock market would at least yield 10 percent annual growth in the average scenario, in addition to dividends. One of the greatest blockers for individuals to enter the Saudi stock market is the uncertainty of the investor about the market and how relevant the news to a specific stock.

Moreover, the perception of stocks isn't always rooted in objective analysis; instead, it's often influenced by sensationalized news headlines and prevailing public sentiment. For many, discerning the direct impact of global events or economic news on specific stocks or the broader market is a challenging endeavor. The rapid pace of information dissemination, combined with the volume of news, can create a cacophony that makes it difficult for the average person to filter out noise from genuinely impactful events. Consequently, potential investors may feel unequipped to make timely and informed decisions, fearing that they might miss crucial cues or misinterpret signals, leading them to either make impulsive choices or avoid the stock market altogether.

Our approach is to provide users, potential investors, with the means to understand and invest in the stock market; through stationary news and historic stock price analyses over companies in the Saudi stock market.

2 Model Architecture

Time series forecasting is the process of predicting future values based on previously observed values. Over the years, several methods and technologies have been developed for this purpose. This chapter delves into the prominent technologies and models used for time series forecasting.

Statistical Models

The allure of predicting the stock market has long led to the application of time series forecasting models. ARIMA, which stands for Auto-Regressive Integrated Moving Average and introduced by Box and Jenkins (1970), is one of the most widely used time series forecasting methods. The model is defined by three primary parameters: p (autoregressive term), d (differencing term), and q (moving average term). The mathematical representation of ARIMA is:

$$(1 - \Phi_p B)(1 - B)^d Y_t = (1 + \Theta_q B)\epsilon_t \tag{1}$$

Where Φ and Θ are the parameters of the AR(p) and MA(q) models respectively, B is the back shift operator, and ϵ is white noise [3]. Its core lies in the autoregression (AR) component, represented by the equation:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t \tag{2}$$

Exponential smoothing models are used for forecasting data with a trend and/or seasonality. The most common of these models is the Holt-Winters method, which has three components: level, trend, and seasonality. The equations are:

$$l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \tag{3}$$
$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \tag{4}$$
$$s_t = \gamma(y_t - l_t) + (1 - \gamma)s_{t-m} \tag{5}$$

Where l, b, and s are the level, trend, and seasonality estimates at time t, respectively, and α, β, γ are the smoothing parameters [4].

This method is closely related to STL, Seasonal Decomposition of Time Series by Loess; a method by Cleveland et al. (1990) [5], decomposes a time series into seasonal, trend, and residual components, expressed as:

$$Y_t = S_t + T_t + R_t \tag{6}$$

Deep Learning Models

Convolutional Neural Networks (CNN), originally designed for image processing, have been adapted for time series forecasting, especially when the patterns in the series can be spatially characterized (LeCun et al., 1998) [6]. Characterized by convolution operations, are given by:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t - a) \tag{7}$$

Long-Short Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is particularly

suited for time series forecasting due to its ability to capture long-term dependencies. It uses three main gates: input, forget, and output, which control the flow of information through the network. Given an input series, an LSTM updates its state using the following equations:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{8}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{9}$$
$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{10}$$
$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{11}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{12}$$
$$h_t = o_t \times \tanh(C_t) \tag{13}$$

Where $f, i, \tilde{C}, C, o, h$ represent the forget gate, input gate, cell input, cell state, output gate, and hidden state respectively [7].

Transformer-based Models, such as the BERT and its variants, have mechanisms like self-attention that can weigh the significance of different time steps in the series, thereby potentially outperforming RNNs and LSTMs in certain applications (Vaswani et al., 2017) [8]. For Transformer-based models, the self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

Hybrid Models

ARIMA-LSTM, combining the statistical prowess of ARIMA and the deep learning capabilities of LSTM, this model captures the underlying patterns and long-term dependencies in the data (Bao et al., 2017) [9]. Wavelet-CNN approach combines the wavelet transform with CNNs. The wavelet transform helps decompose the time series, extracting essential features, which are then passed to CNNs for forecasting (Borovykh et al., 2017) [10], with wavelets described by:

$$W_f(s, \tau) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-\tau}{s}\right) dt \tag{15}$$

2.1 Our Model Structure

In this project, we are leveraging the use of LSTM-based model for time-series forecasting, with the integration of a pre-trained Transformer-based model, namely AraBERT Base-v2 [11], for sentiment analysis to correlate and translate the financial news to the LSTM model.

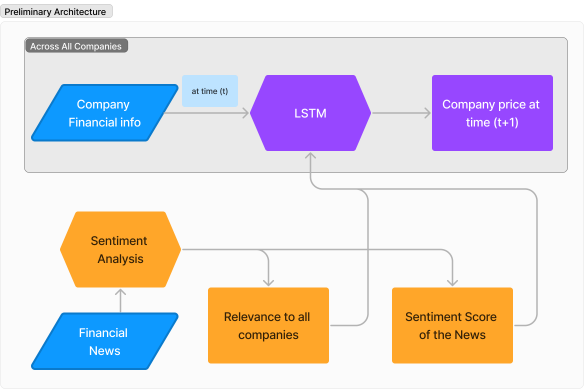

Figure 1: The architecture of the multi-model structure.

The architecture is proposed in Figure 1, encapsulates the financial news, with a provided sentiment score from the sentiment analyzer, which would then be propagated to the LSTM-based model to use as a concatenated input with the financial information. Which includes the stock information of each company: Open price, High, Low, Volume of transactions; and other econometric measures, including: Inflation rate, Unemployment rate, local GDP, stock market indicators (TASi and SP500), Oil price change, Gold price, and REPO rate. The current strcuture of the LSTM-based model is as follows: three stacked LSTM layers followed by 2 dense layers to a final single output.

3 Model Building

The process of building the model consisted of four main steps, collection and preprocessing of data followed by training and hyperparameters fine-tuning.

3.1 Data Collection

The model had two main types of data to be collected: numerical information about the companies and economical environment and textual information about the financial news attributed to each company.

Furthermore, we started collecting the numerical information first. The process was fairly straightforward, collecting information in a systematic manner was as followed:

- Collect stock information of each company from Yahoo finance and Tadawul websties [12, 13].
- Collect information about the econometric measures of the market from the Ministry of Statistics [14].

Collecting stock information from Yahoo finance and Tadawul was done by leveraging a scraping script that accesses the Yahoo website and collect information via Here is an inline code snippet:

```
from pandas_datareader.data import DataReader
import yfinance as yf
from pandas_datareader import data as pdr
```

Listing 1: Yahoo finance stock information collecting

By simply providing the company code-name (e.g. 2222.SR for Aramco), the information is collected from the specified range of dates. The codes were collected from Tadawul website via simple data scraping snippet using *Selenium* library.

As for the textual information, this was collected over two steps: collected general financial news to establish base knowledge and Saudi-market financial news to be domain-specific. The first was provided via an open source dataset, Financial Phrase Bank on huggingface [15, 16]. The latter was collected from a local website, Mubasher, using the *Selenium* library for scraping [17].

	General Knowledge	**Domain-Specific Knowledge**
Size	4910	5454
Source	Financial Phrase Bank	Mubasher

Table 1: The data used for training the sentiment model

The domain-specific data was not labeled, and in this project the data was manually labeled, following the same labeling style done in the Financial Phrase Bank, all news are either negative, neutral, or positive; numerically: 0, 1, or 2, respectively.

3.2 Data Preprocessing

For textual data, all data-points were embedded into a pre-trained Transformer-based model, namely AraBERT Base-v2 [11], for tokenization and preparing the data for training. Tokenization splits the data into tokens; segments of sub-words in a series format, before translating them into unique identifiers; embedding.

For numerical data, all data-points across the same company are normalized, in order to have a multi-company model prediction. The normalization is done in batches; to limit the learned mean and standard deviation in each batch, to eliminate any *future peaking* done by the model.

3.3 Model Training and Finetuning

The two models were trained independently, such that, the sentiment model would be used to make the predictions over the news and pass the information to the time-series model, namely our LSTM-based model, to make the prediction of the price.

The models were trained for 100 epochs and the following specifications were given to each model:

- Layer sizes of powers of 2.
- Learning rate of 0.001.
- Batch size of 16.
- Optimizer Adam.
- Mean Absolute Error, for the time-series.
- Cross Entropy, for the sentiment analysis.

The sentiment analysis model achieved a score of 68 percent as shown in 2, which is potentially attributed to the lack of sufficient data and the possibility of mislabeling the news data. The time-series

	precision	recall	f1-score	support
0	0.54	0.59	0.57	73
1	0.76	0.76	0.76	306
2	0.61	0.59	0.60	169
accuracy			0.68	548
macro avg	0.64	0.65	0.64	548
weighted avg	0.68	0.68	0.68	548

Figure 2: The architecture of the multi-model structure.

model achieved a mean-absolute-error value of 1.038 riyal, over Rajhi stock price. While this result is very promising, the time-series data might behave differently in the performance due to variability in distributions, which entails another level of complexity.

4 Discussion

In the realm of natural language processing, the choice of model architecture significantly influences the performance, complexity, and interpretability of the underlying model. LSTM (Long Short-Term Memory) models have long been a popular choice due to their ability to capture sequential dependencies. However, when considering higher complexities and the requirement for explainability, LSTM models might not always be the optimal choice. LSTMs are adept at capturing local dependencies and short-term patterns in sequential data. However, their performance tends to plateau when faced with more intricate and intricate linguistic structures, such as long-range dependencies and complex syntax. This limitation can become apparent in tasks involving intricate language nuances or broad context comprehension. The inherent design of LSTMs, with their recurrent connections, might not allow them to capture such complexities effectively. Furthermore, the interpretability of LSTM models might be limited. Due to their sequential nature, it can be challenging to discern the exact decisions or features that contribute to the model's predictions. This opacity could hinder the ability to understand why certain choices are made or identify potential biases within the model. The emergence of transformer-based models, represented by the likes of BERT and GPT variants, ushers in a paradigm shift. Through intricate attention mechanisms, these models transcend rigidity and embrace flexibility. This enables them to not only grasp extensive global dependencies but also shine in deciphering convoluted language structures and spanning long-range connections. The allure of transformer models magnifies when considering their potential for conditional actions and customizable adaptations. Furthermore, transformers provide an edge in terms of explainability. The attention mechanisms can be visualized to highlight which parts of the input data contribute most to specific predictions, offering insights into the model's decision-making process. This level of transparency is invaluable, especially in critical applications where model behavior needs to be understood and scrutinized. Another notable advantage of transformer models is their capacity for conditional actions and customization. Transformers can be fine-tuned for specific tasks while retaining their pre-trained knowledge. This enables fine-grained adjustments to cater to different problem domains, striking a balance between general language understanding and task-specific nuances.

5 Conclusion

In conclusion, the use of transformer-based models offered a great encapsulation of meaning to the news and improved the performance of the time-series model. However, due to low amount of data-points and manual labeling, potential mislabeling is higher than usual, as professional labeling was not done at this stage; and the need of more data-points to have a better performing model. Additionally, while LSTM models have their merits in certain scenarios, they might not adequately address higher complexities and interpretability demands. Transformer-based models might be a more flexible, adaptable, and explainable solution. By effectively capturing global dependencies, enabling transparency through attention mechanisms, and allowing for customization, transformer models pave the way for improved performance and better model understanding in the context of complex language tasks.

6 Future Work

While this project has merits, more work is to be investigated. This project will continue to tackle the challenge by :

1. Improving the Sentiment Analysis model would be essential and over two main steps:
    1.1. Getting more data for the domain-specific headlines to saturate the model with enough data.
    1.2. Converting the model structure and data to produce a sentiment score rather than a class label, which would account for more complex data.
2. Improving the stock prediction model, namely the LSTM-based model, would also be crucial and over three main steps:
    2.1. Exploring the explainability and feasibility of using Transformer based model, Temporal Fusion Transformer. [18]
    2.2. Introduce hierarchical structure in our model that would encapsulate multiple vertical levels i.e., for more compatible sentiments with sectors and companies.
    2.3. Incorporate financial statements within the model data for each company.
    2.4. Highlight factors of increase and decrease of a stock price, based on all input values from all models, stationary and non-stationary data; explain the behavior.

References

[1] The progress achievements of saudi arabia - vision 2030. Accessed on: Fri, 25 Aug 2023.

[2] Jeremy J Siegel. *Stocks for the long run: The definitive guide to financial market returns & long-term investment strategies*. McGraw-Hill Education, 2021.

[3] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel. *Time series analysis, forecasting and control*. Holden-Day, 1976.

[4] Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Research Memorandum*, 1957.

[5] Robert B Cleveland et al. Stl: A seasonal-trend decomposition. *Journal of Official Statistics*, 1990.

[6] Yann LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[8] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[9] Wei Bao et al. A hybrid model for time series forecasting. *Journal of Machine Learning Research*, 2017.

[10] Anna Borovykh et al. Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*, 2017.

[11] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*, 2020.

[12] Saudi stock exchange (tadawul). `^5^`. Accessed: August 25, 2023.

[13] Yahoo finance - stock market live, quotes, business & finance news. `^1^`. Accessed: August 25, 2023.

[14] Ministry of Statistics, Saudi Arabia. `https://www.stats.gov.sa`. Accessed: August 25, 2023.

[15] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65, 2014.

[16] Hugging Face Inc. Hugging Face. `https://huggingface.co`. Accessed: August 25, 2023.

[17] Mubasher Info. Saudi stock exchange, 2023. Accessed: August 21, 2023.

[18] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.