Arabic Financial Stock Prediction Approach Based On Deep Learning Technique

Maryam Mohamad Al-Zawahra, Mohammad Bany Baker, Mohammad Zayed Almuiet

Abstract: The prediction of the financial stock market based on the published financial news is considered to be a major issue in the financial field, particularly in light of the significant contents of the news article. Nevertheless, the prediction technique used to gather useful terms from the Arabic financial news has been elusive and is yet to be established. Therefore, this study aimed to develop and propose a practical solution by which the Arabic stock market can be predicted through the analysis of event terms from the Arabic financial news articles. The approach obtained 89.68% accuracy in such prediction, indicating the usefulness of the approach. This is a unique and pioneering study that applies machine learning method to predict Arabic stock market. The study has several implications to the financial field and profession.

Index Terms: Financial stock market, Prediction, Machine learning, Deep Learning, Recurrent Neural Network.

1 INTRODUCTION

Stock market is a crucial component of the financial market, in terms of the interactions and dealings that take place through it. Investor and speculator circles in the market are looking to raise profitability by the analysis of the information provided concerning the market. In this regard, one of the main market information sources is the financial news articles and such articles have been extensively utilized for analysis among investors [1]. In the current era of data, referred commonly as the Big Data Era, news articles have shown a monumental increase, and faced with a considerable pile of news articles, institutions have been increasingly turning to high processing computer power for analysis. Support systems predictions can help investors sift through noises and reach wise and informed decisions as contended by [2]. Therefore, this study proposes a model to conduct news articles analysis for accurate predictions to solve the issue. Moreover, predictive measures can basically be divided into two major analyses, which are technical and fundamental analysis. They are distinct from each other according to the data input, and also, the former is used with historic market data, while the latter is used for news concerning country, society, firm, and the like. Majority of prior literature has conducted technical analyses based on quantitative historic market data and the traders’ inclination towards the use of technical qualitative techniques. In relation to this, basic data is more difficult to use as the data input, particularly if it lacks structure. However, basic data may be obtained from structured and numerical sources such as macro-economic data or regular financial reports provided by the governments and their banks. Such basic data aspect has not been extensively studied, although authors do rarely illustrate their predictive value (e.g., [3]).

Basic data in the form of unstructured text is the most difficult aspect to examine and is thus, the direction that the study adopted. This may be exemplified by the textual information available in social media, blogs, forums and news [3]. In the field of technical analysis, researchers have primarily made use of mathematics for the analysis of historical stock price patterns, and eventually predict stock prices for the future. Many algorithms have been adopted for such purpose in different forms; for instance, multiple kernel learning in [4], deep learning in [5] and [6], stepwise regression analysis in [7], among others. Despite the clear and consistent outcomes achieved, it remains challenging to predict stock prices in an accurate manner through the sole use of historical prices owing to the unpredictable events that can influence such prices. Meanwhile, in [8] study, the authors employed fundamental analysis of natural language processing (NLP) in their attempt to analyze financial news and statements from the firm, with the ultimate aim of predicting future stock trends (uptrend and downtrend). Specifically, NLP’s bag-of-words technique is the most often used method for the features extraction from the news articles, measuring each word’s presence and transforming text information into vector spaces with their help. Following this step, machine learning algorithms are used to establish the relationship between word patterns and the movements of stock prices. Regardless of the accurate ability of bag-of-words technique in several studies (e.g., [5], [8]), one crucial element has been left out in the prediction of directions and that is the sentiment of the article. In relation to this, a specific phase in concerns the interpretation of the investors of the published news articles after which they are transformed into positive/negative sentiments to reach decisions as to whether stocks should be purchased, held or sold on the basis of the interpretations of sentiments. Lastly, the aggregate of the market prices is used on each investor’s actions, after which, this is represented in the final price trends. In other words, based on the above, the combination of sentiment analysis and natural language processing could prove effective. Lastly, machine learning algorithms are used to obtain information on the relationship between extracted news features and stocks trends. More recently, Deep Neural Networks (DNNs) – a category of machine learning – has garnered success after another in different fields, for instance, in speech recognition in [9] and computer vision in [10]. In this line of study, neural networks have become popular owing to the image-based applications
and this holds true for the Recurrent Neural Networks (RNNs)—image-based applications. They have been utilized, with success, in the prediction of future events in time series and have contributed to developing various fields centered on the prediction of events sequence. However, owing to DNN models efficiency in handling large-scale data sets [11], several studies have adopted them on features extracted from news articles and historical prices (e.g., [12]). In the present work, the authors propose a new approach known as the Arabic Financial Stock Prediction (AFS), with the included construct being Arabic Financial Lexicon (AFL), which stores event terms obtained from a year data set. The approach is proposed to predict specific directions of stock price based on the articles from the news (up or down) at the day’s end. The work aims to contribute to providing an efficient predictive model to enhance the accuracy of the semantic-based stock price movement prediction.

2 LITERATURE REVIEW
Owing to the inability of machine learning algorithms to process raw text/sentiment representations, an additional representation layer is required, and one major method used is to represent article terms in binary terms (whether or not the term is present in a specific article) [13]. Such a binary solution is primarily used with the help of a matrix, within which the articles constitute an axis, while the terms in the corpora and other features constitute another axis. This leads to large but sparse matrices, with the terms through the dataset outnumbering those in individual articles [14]. After the representation of the financial news articles, learning algorithms are used for the identification of behavioral patterns upon which the model predictions are based on. Support Vector Regression (SVR) is one common accepted method akin to Support Vector Machines (SVM), sans the classification feature [15]. Similar to SVM, SVR mitigates the error in the fit while increasing the goal function by fitting a regression estimate via a multidimensional hyperplane. The method is suitable to be used in addressing textual input in binary process, particularly in similar financial news studies [16]. In recent times, there is higher likelihood to employ deep learning algorithms in stock market prediction as opposed to using economic models, as evidenced by studies on several deep learning models, supported by measured outcomes to indicate the optimum use of models. For instance, [17] developed a model for the stock market price prediction with the help of three neural network algorithms, namely, the Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), with the analyzed data obtained from Tata Motors, Tech Mahindra, and TVS Motors in the automobile firms. Meanwhile, the firms included in the banking sector included Syndicate bank, which is used for the purpose of training and was tested on three banks (Union Bank, Yes Bank and Syndicate Bank). The study conducted a comparison between three algorithms and indicated LSTM to be the most optimum model, producing less error percentage in both the automobile and banking sectors. Moreover, in [18] study, the LSTM was used with an integrated ELSTM layer, and data was represented in stock vector. Based on the findings, the model accuracy obtained a 53.2%, with 57% A-share Composite Index. Similarly, [19] brought forward an integrated Particle Swarm Optimization (PSO) model with least square support vector machine (LS-SVM) (POS-LS-SVM) after which they compared it to LS-SVM and ANN-BP. They found the proposed model to have the prediction of the lowest error. Stock market prediction was also attempted in [20] study, with RNN approach, and input represented in character-level language model rather than word-level language (characterized by inaccuracy in rare words), sentence embedding input and other complex input models. Based on their findings, the model managed to provide 63% inter day, 59% intraday of S&P 500 Index prediction, with 64% inter day and 61% intraday of individual stock prediction. The model was found to be simple, promising and competitive compared to its counterpart models. In the same line of study, [21] made use of Optimal Long Short-Term Memory (O-LSTM) deep learning and adaptive Stock Technical Indicators (STIs), on three datasets. They showed the model to obtain accurate results compared to SVM, LR and ELSTM. Meanwhile, the effects of twitter on stock market predictions were the focus of [22] study of market predictions. More specifically, they made use of financial tweets to compute sentiments and volume for the indices in Saudi Arabia, Kuwait, Dubai, Abu Dhabi, Qatar and Egypt. Based on the finding of the Granger causality, the tweets-financial market and Pearson correlation relationships showed accurate correlations, particularly between twitter sentiment and volume and change of stock market in the context of the index in Egypt. In the US stock market, [23] brought forth a model to predict next day US stock markets, employing SVM algorithm and comparing the accuracy results with those of MART algorithm on the financial data provided by NASDAQ, DJIA, S&P 500. The findings indicated 74.4%, 77.6% and 76% accuracy on the financial data respectively, with mean square error (RMSEA) calculated for SVM and other algorithms and found to be 21.6. Furthermore, [24] indicated that SVM algorithm and Radial Basis Function (RBF) were able to show efficient and accurate predictions compared to other models. Lastly, four types of deep learning were used in [25] and they were Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for the prediction of daily price stocks of the Bank of America and Chesapeake Energy NYSE. They found CNN model to provide the best result compared to others.

3 PROPOSED APPROACH
This study mainly focuses on the closing price of the top 100 stocks in the Stock Exchange of Jordan (Dataset1) as well as on the online financial news articles. Efficiency of the proposed approach was measured using 3 categories of experimental datasets – in the first one, the proposed Arabic lexicon (financial news articles of over a year period, Feb. 2016-Jan. 2017) was generated; the second one consisted of news articles over a year period (Feb 2017-Jan. 2018), used for the analysis of the event term's efficiency and evaluation of the proposed model and the final one, consisted of news articles over a year period (Feb. 2018 – Jan. 2019), to evaluate the model. According to the obtained results, the proposed model achieved 89.68 accuracy in predicting the movement of stock price among financial firms.

3.1 Arabic Prediction Approach (APA)
The study proposes an AFSP that consists of 4 main stages namely, data pre-processing, event terms analysis, AFL
generation, RNN for prediction, with classifier to take input and predict the rise/fall of the price of the certain period, and model evaluation (refer to Fig. 1).

![Fig. 1. Diagram of Item Parameter (after transformation)](image)

### 3.2 Data Preparation
This step involved the processing of Arabic financial news articles and data from stock price. The following stages were undergone:

i. **Text Pre-Processing**

Under this stage, the words were segmented on the choice of news articles, gathered from https://ase.com.jo over a 3-year period (Feb. 2016 – Jan. 2019). The gathered articles were documented with their affected dates and the trading dates within which the stock prices are affected by the news. The article was exposed to word segmentation process with the help of Arabic dictionary (Cambridge)[26]. The first set of Arabic lexicon, containing meaningful terms that are significant to the stock price directions are gathered, with each term marked as 'entity' or 'predicate'. The former is described as a noun reflecting the topic of an event (bank, profit, income), while the latter is described as a term providing additional information on the former (verb or adjective). Additionally, the terms synonyms are kept in the lexicon. Some terms examples in the initial lexicon set are tabulated in Table 1.

**Table 1**

<table>
<thead>
<tr>
<th>Terms</th>
<th>Term types</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>التكلفة (Cost)</td>
<td>Entity term</td>
<td>تكلفة (Cost)</td>
</tr>
<tr>
<td>-production (Cost)</td>
<td>Entity term</td>
<td>قيمة (Cost)</td>
</tr>
<tr>
<td>الربح (Profit)</td>
<td>Entity term</td>
<td>المكسب (Profit)</td>
</tr>
<tr>
<td>نتائج الربح (Net profit)</td>
<td>Entity term</td>
<td>الربح (Profit)</td>
</tr>
<tr>
<td>الفائدة (Increased)</td>
<td>Predicate term</td>
<td>نقص (Decreased)</td>
</tr>
<tr>
<td>الثمن (Dropped)</td>
<td>Predicate term</td>
<td>هبط (Dropped)</td>
</tr>
</tbody>
</table>

iii. **Stock Price Processing**

In order to predict stock market direction, some studies made use of closing price (e.g. [27], [28]). Thus, each stock price rate of change (ROC) during the period of the news articles were calculated with the use of each stock’s closing price. Therefore, the ROC of a stock i(Si) during trading date j(dj) was obtained through the following equation:

\[
[ROC]_{(Si,dj)} = \left(\frac{C_{(Si,dj)} - C_{(Si,dj-1)}}{C_{(Si,dj)}}\right) \cdot \frac{1}{(dj)}
\]

In the above equation, C(Si,dj) denotes the closing price of stock i on trading date j and P(Si,dj) denotes the closing price of stock i on the prior trading day.

iv. **Financial News Article Labeling**

Under this stage, the news articles terms that are significant to stock i and released during trading date j were labeled as up/down through ROC considerations that are as follows:

a. If ROC(Si,dj) is more than 0, then the stock’s change status is labeled “Up”

b. If ROC(Si,dj) is less than 0, then the stock’s change status is labeled “Down”

c. If ROC (Si,dj) is equal to 0, then the entire news articles significant to stock i during trading date j are dropped from the experiment.

The data pre-processing system results is news corpus and this was used as the input data in the remaining steps (AFL generation, effective event term analysis and model evaluation).

v. **AFL Generation**

The generation of AFL involves the weighing of event term, calculating the total weight and conducting the probability evaluation. Arabic financial news articles gathered for a year (Feb. 2016 – Jan 2017) were used to generate AFL, and the study considered the articles’ titles and contents.
vi. **Recurrent Neural Network (RNN)**

In this phase, Arabic financial news articles were gathered over a year period (March 2016 to Feb. 2017), after which, similar to the lexicon generation system, both the title and contents of the articles were considered. The data set comprised of 3500 specific news articles, with closing prices of 150 stock symbols used as input data. The predictive performance of the proposed method was measured and the RNN employed with high efficiency following the steps below.

#### a. Stock Market Direction Prediction

This step involves the assignment of the news articles with a directional probability on the basis of event terms/similar terms that were found in the news with the help of probability value in AFL. For instance, if \( a_v \) is a news article, then \( [SC] \) \( (a_v,S_i)^U \) and \( [SC] \) \( (a_v,S_i)^D \) are the total scores of stock \( i \) that are “Up” or “Down” respectively, based on the occurrence in the news article \( v(a_v) \). Thus, the directional scores of every news article \( v \) in relation to a specific stock \( i \) that contains \( m \) event terms can be calculated using the following equation:

\[
[SC] \ (a_v,S_i)^U = \prod (k = 1)^m P(t,k)^U \\
[SC] \ (a_v,S_i)^D = \prod (k = 1)^m P(t,k)^D
\]  

(2) \hspace{1cm} (3)

In the above equations, \( m \) denotes the total number of event terms in a news article \( v \), \( P(t,k)^U \) denotes the probability of increase in the stock price, brought on by the event term \( k \). In the probability of the stock price dropping owing to the event term \( k \), is denoted by \( P(t,k)^D \). If \( P(a_v,S_i)^U \) and \( P(a_v,S_i)^D \) denote the probabilities of stock \( i \) “up” and “down” respectively, based on the news articles \( v \) occurrence, then \( P(a_v,S_i)^U \) and \( P(a_v,S_i)^D \) can be obtained using the above equations (2 and 3) through the normalization technique represented by the following equations:

\[
P(a_v,S_i)^U = \frac{[SC] \ (a_v,S_i)^U}{[SC] \ (a_v,S_i)^U + [SC] \ (a_v,S_i)^D} \\
P(a_v,S_i)^D = \frac{[SC] \ (a_v,S_i)^D}{[SC] \ (a_v,S_i)^U + [SC] \ (a_v,S_i)^D}
\]  

(4) \hspace{1cm} (5)

In the above equations, \( [SC] \) \( (a_v,S_i)^U \) and \( [SC] \) \( (a_v,S_i)^D \) denote the total directional scores of news article \( v \) that is linked to stock \( i \), calculated from equations 2 and 3 respectively. Two results can be obtained from such probability terms in a news article \( v \), \( P(a_v,S_i)^U \) denotes the probability of stock price increasing based on the appearance of the content in news article \( v \). Contrastingly, let \( P(a_v,S_i)^U \) > \( P(a_v,S_i)^D \) then, stock \( i \) has a likelihood of increasing based on the appearance of the content in news article \( v \).

#### b. Accuracy Calculation

In order to obtain the proposed approach accuracy, the confusion matrix is used, particularly the following equation:

\[
\text{accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]  

(6)

From the above equation, true positive (TP) denotes the number of times that AFSP correctly predicted the increase in stock price, while true negative (TN) denotes the number of times that it correctly predicted the decrease in stock price. Meanwhile, false positive (FP) denotes the number of times that PLSP predicted the increase of stock price when it reality, it decreased, whereas, false negative (FN) denotes the number of times that it predicted the decrease of stock price when it actually increased. The financial news articles terms possess different impact levels on the stock market price; for instance, a few events can lead to stock market fluctuation, while others have minimal impact on it. Therefore, additional steps were adopted to sift through some event terms in AFL that had minimal influence on the movement of the stock price. Consequently, 530 event terms in the AFL met the condition. The set of event terms showed up in 3600 news articles, and the proposed approach managed to successfully achieve 89.68% prediction accuracy (refer to Table 4).

### 4 CONCLUSION

The RNN model was used in the present study to predict Arabic stock market based on the Arabic financial news articles. To the best of the authors’ knowledge, this is the first study to implement the RNN on this type of information using the machine learning method. The method’s prediction is efficient. The study also illustrated the successful analysis of event terms with a big training dataset, which supports several implications to the financial field for the method’s practicality. Future studies of this caliber are suggested to apply the method on similar tasks.

### 5 REFERENCES


