This section introduce the ***Related Work*** for the stock market analysis using Time Series Analysis and its prediction methods namely Autoregressive Integrated Moving Average, or Auto-ARIMA and Long Short-Term Memory (LSTM).

**3.1 Time Series Analysis**

Time series analysis is a [statistical technique](http://www.statisticssolutions.com/directory-of-statistical-analyses/) that deals with time series data, or trend analysis.  Time series data means that data is in a series of particular time periods or intervals. It is widely recommended as a supplement to analysing behaviour change in reversal or multiple‐baseline experiments [[[1]](#footnote-1)]. The method can be used to identify three kinds of statistically significant behaviour change: (a ) changes in score levels from one experimental phase to another, (b ) reliable upward or downward trends in scores, and (c ) changes in trends between phases. The detection of, and reliance on, serial dependency (autocorrelation among temporally adjacent scores) in single‐subject behavioural scores is emphasized.

Time series analysis has many different objectives for instance forecasting future values of the series, extracting a signal hidden in noisy data, discovering the mechanism by which the data are generated, simulating independent realizations of the series to see how it might behave in the future therefore to estimate the probability of extreme events and eliminating the [seasonal component](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/seasonal-component) from data sets in order to reveal more clearly the underlying trend. Time series analysis usually begins with an attempt to find a mathematical model which provides a good representation of the observed data [[[2]](#footnote-2)].

Forecasting using a time-series analysis consists of the use of a model to forecast future events based on known past events. The model model generally reflects the fact that observations close together in time will be more closely related than observations further apart. Three broad classes of time-series models of practical importance are the autoregressive (AR) models, the integrated (I) models, and the moving average (MA) models. There are models, such as the [autoregressive moving average](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/arma-model) (ARMA) and [autoregressive integrated moving average](https://www.sciencedirect.com/topics/mathematics/autoregressive-integrated-moving-average) (ARIMA) that are combinations of the above three [[[3]](#footnote-3)]. According to the research developed in this field, we can classify the techniques used to solve the stock market prediction problems to twofold.

The first category of related work is*econometric models*, which includes classical econometric models for forecasting. Common methods are the autoregressive method (AR), the moving average model (MA), the autoregressive moving average model (ARMA), and the autoregressive integrated moving average (ARIMA). Roughly speaking, these models take each new signal as a noisy linear combination of the last few signals and independent noise terms. However, most of them rely on some strong assumptions with respect to the noise terms such as *-distribution* and loss functions, while real financial data may not fully satisfy these assumptions.

The second category involves*soft computing based models*. Soft computing is a term that covers artificial intelligence which mimics biological processes. These techniques include artificial neural networks (ANN) , fuzzy logic, support vector machines (SVM), particle swarm optimization (PSO), and many others. Nonetheless this computing based model is not covered in this project.

However, to our knowledge, most of these methods require expertise to impose specific restrictions on the input variables, such as combining related stocks together as entry data, inputting different index data to different layers of the deep neural network, and converting news text into structured representation as input. In contrast, our proposed forecasting model directly uses the historical data collected from the market reports as input, which reduce some serious outliers or noise data in the analysis.

**3.2 Autoregressive Integrated Moving Average (Auto-ARIMA)**

An autoregressive integrated moving average, or Auto-ARIMA, is a statistical analysis model that uses [time series data](https://www.investopedia.com/terms/t/timeseries.asp) to either better understand the data set or to predict future trends.

It is a form of [regression analysis](https://www.investopedia.com/terms/r/regression.asp) that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values. An Auto-ARIMA model can be understood by outlining each of its components as follows [[[4]](#footnote-4)]:-

1. [Autoregression (AR)](https://www.investopedia.com/terms/a/autoregressive.asp)refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
2. *Integrated (I)*represents the differencing of raw observations to allow for the time series to become stationary, i.e., data values are replaced by the difference between the data values and the previous values.
3. [Moving average (MA)](https://www.investopedia.com/terms/m/movingaverage.asp)incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each component functions as a parameter with a standard notation. For Auto-ARIMA model, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

*p*: the number of lag observations in the model; also known as the lag order.

*d*: the number of times that the raw observations are differenced; also known as the degree of differencing.

q: the size of the moving average window; also known as the order of the moving average.

In a [linear regression](https://www.investopedia.com/terms/n/nonlinear-regression.asp) model, for example, the number and type of terms are included. A 0 value, which can be used as a parameter, would mean that particular component should not be used in the model. This way, the ARIMA model can be constructed to perform the function of an ARMA model, or even simple AR, I, or MA models.

*Auto-ARIMA and Stationarity*

In an Auto-ARIMA model, the data are differenced in order to make it stationary. A model that shows stationarity is one that shows there is constancy to the data over time. Most economic and market data show trends, so the purpose of differencing is to remove any trends or seasonal structures.

[Seasonality](https://www.investopedia.com/terms/s/seasonality.asp), or when data show regular and predictable patterns that repeat over a calendar year, could negatively affect the regression model. If a trend appears and stationarity is not evident, many of the computations throughout the process cannot be made with great efficacy.

**3.3 Long Short-Term Memory (LSTM)**

Long short-term memory (LSTM) is an artificial [recurrent neural network](https://en.wikipedia.org/wiki/Recurrent_neural_network) (RNN) architecture used in the field of [deep learning](https://en.wikipedia.org/wiki/Deep_learning) [[[5]](#footnote-5)]. As it has feedback connections, it can not only process single data points (such as images), but also entire sequences of data (such as speech or video). It is most applicable to tasks such as unsegmented, connected [handwriting recognition](https://en.wikipedia.org/wiki/Handwriting_recognition), [speech recognition](https://en.wikipedia.org/wiki/Speech_recognition), and anomaly detection in network traffic or IDS's (intrusion detection systems) [[[6]](#footnote-6)].

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell [[[7]](#footnote-7)].

LSTM network is well-suited to [classifying](https://en.wikipedia.org/wiki/Classification_in_machine_learning), [processing](https://en.wikipedia.org/wiki/Computer_data_processing) and [making predictions](https://en.wikipedia.org/wiki/Predict) based on [time series](https://en.wikipedia.org/wiki/Time_series) data, since there can be lags of unknown duration between important events in a time series. It is developed to deal with the [vanishing gradient problem](https://en.wikipedia.org/wiki/Vanishing_gradient_problem) that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs. It is able to solve complex, artificial long-time lag tasks that have never been solved by previous recurrent network algorithms.

1. Richard R.Jones, Russell S.Vaught, Mark Weinrott “Journal of Applied Behaviour Analysis” Volume 10, Issue 1 (Spring 1977) [↑](#footnote-ref-1)
2. P.J.Brockwell “International Encyclopedia of Education; 3rd Edition (2010) [↑](#footnote-ref-2)
3. S. Sinharay “International Encyclopedia of Education; 3rd Edition (2010) [↑](#footnote-ref-3)
4. <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp> (James Chen; 13 April 2019) [↑](#footnote-ref-4)
5. [Sepp Hochreiter](https://en.wikipedia.org/wiki/Sepp_Hochreiter); [Jürgen Schmidhuber](https://en.wikipedia.org/wiki/J%C3%BCrgen_Schmidhuber) (1997). ["Long short-term memory"](https://www.researchgate.net/publication/13853244). [Neural Computation](https://en.wikipedia.org/wiki/Neural_Computation_(journal)). 9 (8): 1735–1780. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1162/neco.1997.9.8.1735](https://doi.org/10.1162%2Fneco.1997.9.8.1735). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276). [↑](#footnote-ref-5)
6. Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; [Schmidhuber, J.](https://en.wikipedia.org/wiki/J%C3%BCrgen_Schmidhuber) (2009). ["A Novel Connectionist System for Improved Unconstrained Handwriting Recognition"](http://www.idsia.ch/~juergen/tpami_2008.pdf) (PDF).  [↑](#footnote-ref-6)
7. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31 (5): 855 -868.  [CiteSeerX](https://en.wikipedia.org/wiki/CiteSeerX_(identifier)) [10.1.1.139.4502](https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.139.4502). [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1109/tpami.2008.137](https://doi.org/10.1109%2Ftpami.2008.137). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [19299860](https://pubmed.ncbi.nlm.nih.gov/19299860). [↑](#footnote-ref-7)