Modeling and prediction of KSE – 100 index closing based on news sentiments: an applications of machine learning model and ARMA (p, q) model

Asma Zaffar¹ · S. M. Aalim Hussain¹

Received: 20 March 2021 / Revised: 2 March 2022 / Accepted: 4 April 2022 / Published online: 18 April 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The main financial markets of every country are stock exchange and consider as an imperative cause for the corporations to increase capital. The novelty of this study to explore machine learning techniques when applied to financial stock market data, and to understand how machine learning algorithms can be applied and compare the result with time series analysis to real lifetime series data and helpful for any investor. Investors are constantly reviewing past pricing history and using it to influence their future investment decisions. The another novelty of this study, using news sentiments, the values will be processed into lists displaying and representing the stock and predicting the future rates to describe the market, and to compare investments, which will help to avoid uncertainty amongst the investors regarding the stock index. Using artificial neural network technique for prediction for KSE 100 index data on closing day. In this regard, six months’ data cycle trained the data and apply the statistical interference using a ARMA (p, q) model to calculate numerical result. The novelty of this study to find the relation between them either they are strongly correlated or not, using machine learning techniques and ARMA (p, q) process to forecast the behavior KSE 100 index cycles. The adequacy of model describes via least values Akaike information criterion (AIC), Bayesian Schwarz information criterion (SIC) and Hannan Quinn information criterion (HIC). Durbin- Watson (DW) test is also applied. DW values (< 2) shows that all cycles are strongly correlated. Most of the KSE-100 index cycles expresses that the appropriate model is ARMA (2,1). Cycle’s 2nd,3rd,4th and 5th shows that ARMA (3,1) is best fitted. Cycle 8th is shows ARMA (1,1) best fit and cycle 12th shows that the most appropriate model is ARMA (4,1). Diagnostic checking tests like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Theil’s U-Statistics are used to predict KSE-100 index cycles. Theil’s U-Statistics demonstrate that each cycle is strongly correlated to previous one.
Keywords  KSE-100 index · ARMA (p · Q) · RMSE · MAE · Theil’s U-statistics

1 Introduction

Stock market forecasting has been at center for years since it can produce major returns. Predicting the closing values of stock market is not an easy task, mainly because of the close to random walk bearing of a stock market time series. It is the pose of trying to examine the future value of a company stock or other scripts traded on an exchange [8]. The golden prediction of a stock’s future closing price could return major profit. Apparently, stock market prediction is categorized into two parts Fundamental Analysis and Technical Analysis.

- Fundamental Analysis wraps analyzing the company’s future profitability on the basis of its current business working environment and financial stability performance.
- Technical Analysis, on the other hand, wraps reading the charts and using statistical figures to forecast the trends and movement in the stock market [5].

The Pakistan Stock Exchange (PSX) has integration of three prevailing stock exchanges from Karachi, Islamabad and Lahore on January, 11th 2016, and developed an incorporated component of Emerging Market Index MSCI, in May 2017. Now, 400 brokerage houses besides 21 assets managing companies are registered members of PSX [20].

The pandemic ruined the limits of China in January 2020 early and provide severe threats to uncontrollable deaths and economies. Due to COVID-19, it is attack all global economies very severely, and directly effects on financial markets which were detected with speedily. The Macroeconomic impacts come with the passage of time appropriately. Overall world experienced significantly decrease 50% up to. Therefore, the long-run and short-run impacts of COVID-19 on economic aspects consideration. Many governments announced US Dollars (multi-trillion) as safety packages and rescue, when its spread started [2].

In the last two decades forecasting of stock returns has become an important field of research. In most of the cases the researchers had attempted to establish a linear relationship between the input macroeconomic variables and the stock returns. But with the discovery of nonlinear trends in the stock market index returns [18], there has been a great shift in the focus of the researchers towards the nonlinear prediction of the stock returns. Although, there after many literatures have come up in nonlinear statistical modeling of the stock returns, most of them required that the nonlinear model be specified before the estimation is done. But for the reason that the stock market return being noisy, uncertain, chaotic and nonlinear in nature. In literature, different sets of input variables are used to predict stock returns. In fact, different input variables are used to predict the same set of stock return data. Some researchers used input data from a single time series where others considered the inclusion of heterogeneous market information and macro-economic variables. Machine learning help to find hidden characteristics in learning process itself. It is a help to find approximation of complex data entry and its output hence machine learning prove that best selection of price for organization. Stock price company’s analysis data very well in deep learning using different strategies. Since so far five countries using this technique namely Hong Kong Canada Japan UK and USA. They trained the data of stock and recover them very well [12].

A recent study [10, 11, 16] has shown that ANN model can be more advantageous compared to other SVM or LR models and the advantages increase in accuracy with multiple
attributes [3]. Works well even if attributes and output do not have a clear relation. Also, some disadvantages must be considered which are brought with time, required for prediction is more than other methods can face overfitting problem.

Machine learning help to find hidden characteristics in learning process itself. It is a help to find approximation of complex data entry and its output hence machine learning prove that best selection of price for organization. Stock price company’s analysis data very well in deep learning using different strategies. Since so far five countries using this technique namely Hong Kong Canada Japan UK and USA. They trained the data of stock and recover them very well. [22]

A statistical approach to forecasting involves stochastic models to predict the values of KSE-100 index by using pervious once within the linear statistic, two methods are frequently employed in literature, viz. Autoregressive AR (p) and Moving Average MA (q) [9, 13]. ARMA models are developed by [21]. An ARMA model is that the mixture of a thought of Autoregressive AR (p) and Moving Average MA (q) process. The concept of ARMA process is strongly relevant in volatility modeling. ARMA model is wieldy used for forecasting the long-term values. Autoregressive process (AR) is developed by [13]. The long range correlation recommends the positive autocorrelations presence that continue significantly high over large time lags, to the autocorrelation function of the series demonstrate a slow asymptotic decay. The persistence or strength of the long-range correlations constrained in experimental time series can be evaluated by various well-known methods [9].

2 Data description and methodology

Stock exchange prediction is one of the best examples of machine learning program since this prediction. Can be validate by statistical tool like testing of data or diagnostic test and furthermore using time series analysis by making a data stochastic hence multivariate technique make our model unique or we can say it is hybrid as most of the papers in stock exchange prediction only uses ARMA (p, q) model which gives good prediction but if we trained the data and validate by test and apply ARIMA (p, d, q) model than it will give the best production which make this article more interesting.

In this study two learning algorithm and two weight initialization methods are compared. The results reported that prediction of stock market is quite possible with both the algorithm and initialization methods but the performance of the efficiency of the back propagation can be increased by conjugate gradient learning and with multiple linear regression weight initialization. In this research, we applied LSTM model to predict the direction of the KSE-100 index stock market closing price.

The feature sets will be followed by:

- News of particular days
- Closing prices and index of previous days
- Market shares up and down,
- Analytics and charting,
- Individual company share predictions as well as market share predictions,
- Company charting and market charting.
The data of KSE-100 index stock from 2015 to 2020 [cycles (1st-12th)] is daily bases under deliberation. Each cycle consists of six-month duration. The data is collected from www.investing.com

The adequate models ARMA (p, q) are selected by Akaike information criterion (AIC), Bayesian Schwarz information criterion (SIC) and Hannan Quinn information criterion (HIC). The forecasting ability of each model of KES-100 index stock cycles will be judged by diagnostic checking tests like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Theil’s U-Statistics. Mean maximum likelihood estimation is used to evaluate ARMA (p, q) model. The Statistical EViews version 9.0 software is used for calculation and analysis of ARMA (p, q) model and respective graphs. For instance, time series plots and fitted, residual and forecasted plots for six-year data of Karachi Stock Exchange 100 Index (KSE-100) on day closing is used which is used in the form of six-month cycles.

This section consists of two subsections. First section covers machine learning techniques description and second section belong to time series analysis techniques.

2.1 Methodology of machine learning

A session describes the machine learning theory and techniques:

2.1.1 Neural networks and machine learning

Rather than straight and determined backslides which are seen as immediate models, the objective of neural associations is to get non-straight models in data by adding layers of limits to the model. In the image underneath, the clear neural net has three information sources, a singular covered layer with five limits, and a yield layer. In all honesty, the development of neural associations is adequately versatile to manufacture our striking immediate and determined backslide. The term deep taking in comes from a neural net with many covered layers and epitomizes a wide arrangement of models. It is especially difficult to remain mindful of upgrades in significant learning, somewhat considering the way that the assessment and industry networks have duplicated down on their significant learning attempts, delivering very surprising strategies reliably. For the best display, significant learning methods require a huge load of data — and a huge load of interaction power since the procedure is self-tuning various limits inside titanic designs. It quickly ends up being clear why significant learning specialists need very unbelievable PCs improved with GPUs (graphical getting ready units). Significant learning strategies have been extremely viable in the domains of vision (picture gathering), text, sound and video. The most notable programming packs for significant learning are Tensorflow and PyTorch.

2.1.2 Natural language processing

An epic level of the world’s data and data is in some sort of human language. Unmistakably, PCs cannot yet totally grasp human substance anyway we can set them up to do certain tasks. For example, we can set up our phones to autocomplete our texts or to address inaccurately spelled words. We can even show a machine to have a direct conversation with a human. Normal Language Processing (NLP) is not an AI methodology accordingly, yet rather for the most part used strategy to design text for AI. Consider gigantic heaps of text documents in an
arrangement of associations (word, online destinations, etc.). By far most of these substance files will be stacked with botches, missing characters and various words that ought to have been filtered through. At this moment, the most well-known pack for dealing with text is NLTK (Natural Language ToolKit), made by researchers at Stanford.

The most un-troublesome way to deal with plan text into a numerical depiction is to calculate the repeat of each word inside each substance record. Consider an organization of entire numbers where each line tends to a book record and each fragment tends to a word. This system depiction of the word frequencies is by and large called Term Frequency Matrix (TFM). Starting there, we can make another standard network depiction of a book chronicle by confining each segment on the grid by a heap of how critical each word is inside the entire corpus of records. We call this method Term Frequency Inverse Document Frequency (TFIDF) and it ordinarily ends up being better for AI endeavors.

2.1.3 Word Embedding’s

TFM and TFIDF are numerical depictions of text records that simply consider repeat and weighted frequencies to address text chronicles. Oddly, word embeddings can get the setting of a word in a chronicle. With the word setting, embeddings can assess the likeness between words, which along these lines grants us to do calculating with words.

Word2Vec is a procedure subject to neural nets that aides’ words in a corpus to a numerical vector. We would then have the option to use these vectors to find reciprocals, perform number shuffling assignments with words, or to address text records (by taking the mean of all the word vectors in a report). For example, we ought to expect that we use a satisfactorily gigantic corpus of text documents to evaluate word embeddings. What about we similarly expect that the words master, sovereign, man, and woman are significant for the corpus. Let say that vector (‘word’) is the numerical vector that tends to the ‘word’. To assess vector (‘woman’), we can play out the number related action with vectors:

Vector (‘king’) + vector (‘woman’) — vector (‘man’) ~ vector (‘queen’).

Word depictions grant finding resemblances between words by preparing the cosine likeness between the vector depictions of two words. The cosine comparability measures the point between two vectors.

We register word embedding is using AI methods, anyway that is consistently a pre-step to applying an AI figuring on top. For instance, expect we approach the tweets of a couple thousand Twitter customers. Furthermore, surmise that we know which of these Twitter customers bought a house. To envision the probability of another Twitter customer buying a house, we can get Word2Vec together with an essential backslide.

You can get ready word embeddings yourself or get a pre-arranged (move learning) set of word vectors. To download pre-arranged word vectors in 157 unmistakable vernaculars, examine Fast Text.

2.1.4 Testing

The testing of the study can be done in two ways:

- Unit testing in which each of the portion of the study will be tested individually.
- SIT (System Integration Testing) in which the whole product will be tested.
For further validation, verification, and satisfaction of the applied working machine learning algorithm the output can be tested by comparing it with the result of SPSS statistical tool.

2.1.5 Evaluation

To evaluate the machine learning models, we will have to know the basic performance metrics of models. We will also be requested to go through other supervised machine learning algorithms.

2.1.6 Approach for sentiment analysis of news

As the stock market is affected with several news and for fetching the daily news of the country condition and stock business market, we have made a news scrapper using python programming language that on executing fetches and stores all the news.

The scrapper is made using requests and the beautiful soup libraries. [12] The news scrapper was made for the scrapping of following official news websites:

- Dawn, [7]
- Daily Times, [6]
- Business Recorder, [4]
- Pakistan Observer, [17]
- The Express Tribune. [19]

The scrapper is used for scraping the news with the corresponding date and all the useful data was scrapped.

For further cleaning and preprocessing of the scrapped news NLP (Natural Language Processing) is used by which sentiment analysis is also done which gives the polarity of the news between 0 and 1 using NLTK (Natural Language Tool Kit Library).

Using NLTK we removed punctuation, stop words removal, lower case all the news, stemming and lemmatization [15]. The positive polarization states that the news would be having a positive trend for the stock market and negative polarity states that the news would be having a negative trend in the stock market whereas the neutral polarity states that the trend would be neutral for the stock market.

2.1.7 Proposed work

**Data Collection** For this study we needed the historical stock market closing prices and the news from different sources.

Including this we fetched the news from 5 different sources i.e., Dawn, Daily Times, Business Recorder, The Express Tribune, Pakistan Observer, using python web scrapper that scrap all the necessary news headlines.

Further all the possible previous year’s data that can fetched from these news websites is fetched for better accuracy. The scrapped news data consist of two columns i.e., News Date and the News Headline.

The historical data of KSE – 100 indexes and other companies’ data is fetched of last 5 years from 1st January 2015 to 31st December 2020. [14]
The historical stocks market data consist of six columns i.e., Open, High, Low, Close, Volume, Value. For the project we just considered Close column for prediction.

**Pre-Processing** The stock market is not easy to understand as the data is unstructured and in raw form. The stock market is also not operating on the weekends and on public holidays. Therefore, a lot of preprocessing must be done on the scrapped data.

We have joined the real news with the stocks data of the same day and have ignored the weekends and public holidays.

The preprocessing of the news is based on label as the research is supervised learning.

For the preprocessing we tokenize the news headlines into words. We removed the noisy words that are not useful for the regression task. Moreover, we dropped the data that only contains numbers, white spaces, tabs, punctuation, stop words, etc. after cleaning all this data we created bag of words relevant to the stocks and the English stop words. After the cleaning of data, we performed stemming and lemmatization for taking the words to their root words such as the words ‘becomes’, ‘becoming’ are taken to its root word ‘become’.

**Input Variables** We get Input variables by doing sentiment analysis on the scrapped news. We used open-source python library of VADER sentiment analysis [17] by using the library we get five factors for the five different news that are the inputs for the prediction which are extracted from real time news. We have joined these input variables with KSE-100 stock market closing price to test and train our LSTM model.

2.1.8 Testing and validation

The 80, 20 Train-Test Ratio was used for the prediction. Training was done on 80% data whereas Testing was done on 20% data.

- X-axis = Date
- Y-axis = Closing Price (both Actual & Predicted)

2.2 Basic Eqs. Of statistical analysis

This section consists of short statistical analysis.

2.2.1 Diagnostic test

The adequacy of selected models is verified by the Akaike information criterion (AIC), Bayesian Schwarz information criterion (SIC) and Hannan Quinn information criterion (HIC). Durbin-Watson Test also apply for knowing most appropriate model. Forecasts with the best-fitted model of KSE-100 index cycles were tested for accuracy with the help of a Root mean square error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Theil’s U-Statistics. A top-level view of these terminologies is given within the subsequent.
**Akaike information criterion** The AIC test was introduced by Hirotogu Akaike in 1973. It is the extension of the utmost likelihood principle the selection criterion is targeting the tiniest amount value of AIC.

\[
AIC = -2 \log \text{(Likelihood)} + 2S
\]  

Where \( S \) is that the model parameter numbers. The chances are that a measure of the fit model. Maximum values exhibit the best fit.

**Schwarz criterion** The SIC test is used to choose out the foremost appropriate model among finite models. The appropriate model relies on the littlest amount value of SIC. Schwarz criterion (SIC) was developed by Gideon E. Schwarz. It’s closely related to the AIC.

\[
SIC = -2 \log \text{(Likelihood)} + (S + S \ln (N))
\]  

Where \( S \) is that the model parameter numbers. \( N \) exhibits the quantity of observations.

**Hannan-Quinn criterion** The HIC is that the criterion for model selection. This test is an alternate to AIC and SIC.

\[
HQC = -2 \log \text{(Likelihood)} + 2 (S + S \ln (N))
\]  

Where \( S \) is that the model parameter numbers. \( N \) exhibits the number of observations.

**Durbin-Watson Test** The DW statistics might be a test for measuring the linear association between the adjacent residual from a regression model. The hypothesis of Durbin-Watson statistics is \( \tau = 0 \) is that the specification.

\[
U_t = \tau U_{t-1} + \epsilon_t
\]  

Durbin-Watson (DW) is adequate 2 shows there is not any serial correlation. If Durbin-Watson (DW) may be a smaller amount than 2 indicate that correlation and thus the range from 2 to 4 represents that negative correlation. The series is strongly correlated if the price nearly approaches to zero.

**Mean Absolute Error** The mean absolute error is expressed as a mathematically formed.

\[
MAE = \frac{1}{n} \sum_{t=1}^{n} |\epsilon_t|
\]  

Where \( n \) is that the number of observations. Mean Absolute Error (MAE) processes absolutely the deviation of forecasted values from real ones. It is also called Mean Absolute Deviation (MAD). It expresses the magnitude of overall error caused by forecasting. MAE does not wipe out the effect of positive and negative errors. MAE does not definite the directions of errors. It should be as small as possible permanently forecasting. MAE depends on the information transformations and thus the dimensions of measurement. Extreme forecast error does not exist by MAE [1].

**Mean Absolute Percentage Error (MAPE)** The Mean Absolute Percentage Error (MAPE) is defined as
Mean Absolute Percentage Error (MAPE) provides the proportion of the everyday absolute error. It is independent of the scale measurement. MAPE does not locate the direction of error. The acute deviation is not penalized by MAPE. During this measure, opposite signed errors do not offset each other in MAPE [8]. This implies that due to the benefits of freedom and commentary on absolutely the share error (MAPE) scale, one in every of the foremost extensively used measures of prediction accuracy. Whereas it is independent of the scale of measurement but full of data transformations (Schwabe H. 1844).

**Root Mean Squared Error (RMSE)** The root mean squared error (RMSE) is defined as

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2}
\]  

RMSE calculates the common squared deviation of forecasted values. The choice signed errors do not offset one another. RMSE provides the entire idea of the error that happened during forecasting. By using the accuracy measures, errors that are small and are getting good, such as 0.1 RMSE and 1% MAPE, can often be achieved. In RMSE, the full forecast error is ill with the large individual error. As an example, an oversized error is way dearer than small errors. It does not reveal the direction of overall errors. RMSE is laid low with the information transformation and the change of scale. RMSE may well be an honest measure of overall forecast error (Adhikari R. 2013).

**Theil’s U-Statistics (U)** Theil’s U-Statistics is defined as

\[
U = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2}} \quad 0 \leq U \leq 1
\]  

Where \(f_i\) represent the forecasted value and \(X_i\) shows that the actual value. U is the normalized measure of the total forecast error. U is equal to 0 exhibits the perfect fit.

### 2.2.2 Tests for normality

The normality test is executed to test whether the data into consideration is usually distributed or not. These tests are supported the analysis of two numerical measures, the shape skewness and also the surplus kurtosis. The information sets are normally distributed if those measures are near zero. The acceptance of Jurque-Bera test also focused on skewness and kurtosis. Hence, the test of normality consists of checking the skewness and kurtosis on which the Jurque-Bera test relies.
Skewness The skewness determines the degree of asymmetry of the information.

\[
Skewness = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{(n-1)S^3}
\]  

(9)

Where \(\bar{X}\) is that the mean and \(S\) is that the variance and \(n\) is that the number of values (Christian and Jean-Michel 2004). The skewness of the normal distribution. If the knowledge is typically distributed, then the skewness shows that the following data is symmetry. If the knowledge is usually distributed if the symmetric distribution (skewness value is capable zero). The distribution is positively skewed, if it is greater than zero and negatively skewed if it’s but zero.

Kurtosis The Kurtosis measures the degree of peakness of the data. Kurtosis has been estimated as

\[
Kurtosis = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{(n-1)S^4}
\]  

(10)

Where \(\bar{X}\) is that the mean, \(S\) is that the variance and \(n\) is that the number of values of the statistical data. Kurtosis of a customary distribution is termed mesokurtic if it is up to 3. Whereas, it is leptokurtic if the value is bigger than 3. It is Platykurtic if the worth is also a smaller amount than 3.

Jurque-Bera Statistics Test (JBS) The JBS is accepted with the normality of the information with skewness is adequate zero and excess kurtosis is additionally up to zero. Jurque-Bera test is defined as follows.

\[
Jurque-Bera\ test = \frac{n(Skewness)^2}{6} + \frac{n(Kurtosis-3)^2}{24}
\]  

(11)

Jurque-Bera test statistics are estimated as Chi-squared distribution with two degrees of freedom. Null hypothesis (HO) could even be a traditional distribution with skewness zero and excess kurtosis zero (which is that the identical as a kurtosis is 3). Alternate hypothesis (HA) of given data is not normally distributed.

2.2.3 ARMA model

In stochastic process, Autoregressive process AR (p) can be expressed by a weighted sum of its previous value and a white noise. The generalized Autoregressive process AR (p) of lag p as follow

\[
X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_p X_{t-p} + \epsilon_t
\]  

(12)

Here \(\epsilon_t\) is white noise with mean \(E(\epsilon_t) = 0\), variance \(\text{Var}(\epsilon_t) = \sigma^2\) and \(\text{Cov}(\epsilon_t, \epsilon_s) = 0\), if s \(\neq\) 0. For every t, suppose that \(\tau_i\) is independent of the \(X_{t-1}, X_{t-2}, \ldots\). \(\tau_i\) is uncorrelated with \(X_t\) for each s < t. AR (p) models regress is past values of the data set. Whereas MA (q) model relates with error terms as a descriptive variable (Hipal et al. 1994). The generalized Moving Average
process MA (q) of lag q as follows.

\[ X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \beta_q \varepsilon_{t-q} \]  
(13)

The process \( X_t \) is defined by the ARMA model

\[ X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \beta_q \varepsilon_{t-q} \]  
(14)

With \( \varepsilon_t \) is an uncorrelated process with mean zero. The prediction of ARMA (p, q) process shows the decay to be sinusoidally and exponentially to zero.

### 3 Results and discussion

User interface will give financial professionals the ability to track and analyze financial data and daily news feed, and provide analytic charts to meet these needs, all on a single platform. Based on previous data the future predictions will be shown via graphs and after giving the news the next day prediction will also be shown daily. The graphs will show KSE-100 stock rates helping the user to understand how the stock market works, making it easier for new investors to join the market and make investments based on trusted data. Most financial data providers are expensive to access and require a monthly or weekly subscription, in which case a new or a small-time investor might not pay for it as it might be of little use to them. Websites will be free of cost, so anyone who wants to learn or invest in the stock market can have access without paying huge amounts.

Figure 1 explains the final train and test validation of KSE 100 index. The green dotted line explains the training of the dataset and the red line shows the prediction. Figure 2 shows Model Loss of 50 epochs of KSE 100 index. The green line indicates the train loss, and the yellow line indicates the test loss. Figure 3 exhibits Points Vs Date of the Actual data of KSE 100 index from the period of 2015–2020. Figure 4 expresses True vs Predicted Value of KSE 100 index. Figure 5 defines Test vs Predicted Value of KSE 100 index. The blue dotted line shows the test validation, and the red line shows the predicted value. Figure 6 shows Future Prediction of six months of KSE 100 index closing. Figure 7 displays Sample of Scrapped News of KSE 100 index (2015–2020) the scrapped news contains the title of every news available on the sites. Figure 8 indicate Sentiment Analysis polarities of Scrapped News (2015–2020) all the sentiments polarities came from the VADER sentiment analyzer library. Figure 9 shows Sample of word count and polarities of Scrapped News (2015–2020) all the word counts are calculated using the word cloud and the polarities are calculated using the sentiment analyzer. Figure 10 present Neural Network LSTM Model used for KSE 100 index Prediction in which six feature sets are used including the closing price and five news for each day and output comes for the predicted closing KSE 100 indexes. Figure 11 expresses the Workflow for the overall prediction of the KSE 100 index. The workflow explains that the raw data comes first then the features are extracted then the data is split into testing and training then the trained data gives the predicted closing index for KSE 100 index. Figure 12 shows Research work working for the prediction of KSE 100 index. It explains that the technical analysis is applied on the historical stock data and
fundamental analysis is applied on the NLP results, prediction prices and the sentiment analysis is applied on the news feeds. Further the trained model is created, and the keyword are extracted from the news feeds from which the predicted values are obtained. Figure 13 shows Sample of Historical Stock Data of KSE 100 index (2015–2020) the data contains Date, Open, High, Low, Close, Volume, value of every day. Figure 14 describe Description of KSE 100 index Stock Data (2015–2020) the data contains the count, mean, standard deviation, min, 25%, 50%, 75%, max. Figure 15 shows Sample of Bag of Words of the Scrapped News (2015–2020). Figure 16 shows Sample of Words Count from the Scrapped News (2015–2020). Figure 17 describe Sample of News Data Preprocessing for the Scrapped News (2015–2020) using NLTK before the polarity detection and after the polarity detection. Figure 18 expresses Sample of News Data Sentiment Polarities for the Scrapped News (2015–2020) these news polarities were taken out from VADER Sentiment analyzer for each day five news. Figure 19 shows Predicted Closing index of KSE 100 from our trained model of LSTM neural network. Table 1 demonstrates the descriptive statistics of each KSE-100 index cycle. Mean, median, standard deviation, skewness, kurtosis and Jurque-
Bera test have been calculated of each cycle. 1st cycle has 128 closing index with 40,498.80± 24.35. 2nd cycle follows 128 closing index with 36,034.35± 4377.675. 3rd cycle expresses 128 closing index with 34,731.29± 3792.565. 4th cycle represents 122 closing index with 37,646.89 ± 2224.751. 5th cycle explores 125 closing index with 40,485.38 ± 1615.475. 6th cycle has 124 closing index with 43,802.92 ± 1354.415. 7th cycle follows 125 closing index with 42,075.09 ± 2318.752. 8th cycle reveals 124 closing index with 49,194.90± 1419.163. 9th cycle represents 125 closing index with 41,415.64± 2393.466. 10th cycle explores 128 closing index with 33,999.88± 38,777.00. 11th cycle expresses 123 closing index with 34,044.59± 1153.430. 12th cycle has 126 closing index with 33,284.96± 1087.421. Each KSE-100 closing index cycle shows that positive skewed (right tail) expect 1st, 4th, 5th and 12th cycles. Whereas different KSE-100 closing index cycles have kurtosis behavior. Maximum KSE-100 closing index cycles shows platykurtic which represents

![Fig. 2 Model Loss of 50 epochs of KSE 100 index](image)

![Fig. 3 Points Vs Date of the Actual data of KSE 100 index](image)
that flat tail (peakness). KSE-100 closing index cycles 1st, 8th, 9th and 12th evaluate leptokurtic (heavy tail). In Jurque-Bera test, KSE-100 closing index cycles each rejected the null hypothesis (sunspot cycles are normally distributed by any mean and variance). The ARMA model is a tool for analysis and calculating of the fundamental structures or getting the predictions of future values in time series. The appropriate models for KSE-100 index cycles are selected based on Durbin-Watson statistics test. The Durbin-Watson (DW) approach to zero is describe the series are positively correlated. The value 2 shows that the series has no correlation. If DW is ranged between 2 and 4 indicates negative correlation. Least square estimation is used to estimate ARMA process. Coefficient of determination, $R^2$ near to one exhibit that data values in each cycle depend on each other. Table 2 depicts the appropriate model for each KSE-100 index cycle. Least square estimation is used to estimate ARMA ($p$,
q) model. Adequacy of the models is checked by AIC, SIC, HQC and Maximum log
lihood tests. Whereas, the forecasting evolution is checked by RMSE, MAE,
MAPE and U tests in Table 3. The criteria of choose most appropriate model is less
values of AIC, SIC, HQC and DW value is less than 2. According to Durbin- Watson
(DW) test and AIC, SIC and HQC represent that the best fitted model is ARMA (2, 1) of cycles 1st, 6th, 7th, 9th, 10th and 11th. Cycles 2nd, 3rd, 4th and 5th are
represented that the most appropriate model is ARMA (3, 1). Cycle 8th expresses that
the best fitted model is ARMA (1, 1) with AIC value 15.2475, SIC value 15.3384 and HQC 15.2844 value. Similarly cycle 12th demonstrate that the most appropriate model is ARMA (4, 1) with AIC value 15.4740, SIC value 15.5641 and HQC 15.5106 value. According to Durbin- Watson (DW) test ARMA (3, 1) model possesses the least value of DW (1.254008) which implies that the 3rd cycle is strongly correlated. Coefficient of determination, $R^2$ demonstrates that data values in each
cycle depend on each other. Cycle 11th exhibits that ARMA (2, 1) is the best model according to log maximum likelihood (−876.0780). Table 3 shows the forecasted evolution of each KSE-100 index cycle. MAPE is least among MAE, RMSE, and MAPE of all the KSE-100 cycles. The smallest values of MAE, RMSE, and MAPE exist for the 3rd KSE-100 index cycle, which is 3516.736, 3126.943 and 0.044968 respectively. Theil’s U-Statistics is also applied evolution of forecasting of each KSE-100 index cycle. According to Theil’s U-Statistics is equal to 0 exhibits the perfect fit. Each KSE-100 cycle has Theil’s U-Statistics value near to zero which is shows that each cycle is correlated to previous one. Theil’s U-Statistics explain that ARMA

![Neural Network LSTM Model used for KSE 100 index Prediction](image1)

![Workflow for the overall prediction of the KSE 100 index](image2)
Fig. 12  Research work working for the prediction of KSE 100 index

Fig. 13  Sample of Historical Stock Data of KSE 100 index (2015–2020)
(1, 1) is best fitted model with 0.014848 value. Figure 20 shows the forecast evaluations RMSE, MAE, MAPE and Theil’s U-Statistics of KSE-100 index of (1st – 12th) cycles.

4 Conclusion

The novelty of this study is a preliminary experiment on how stock market works and is not meant to simulate real-life scenarios but to rather develop a simplistic understanding of finance. It will be beneficial for the investors and brokers to invest as they can check the future predictions, sentiment news, and trend. It will provide the analysis where to invest, including this the predictor will be helpful for all those people who want to invest, even a common man. In this study we have used machine learning techniques to predict KSE-100 stock prices, the machine learning technique used by us is Artificial Neural Networks. We train
the ANN model by using historical stock data. Various features are extracted from the historical stock data. The dataset is then divided into training and testing sets which are used for training and testing the accuracy of the ANN model. The predicted stock prices help investors make smart investment decisions as well as help analysts to predict and study trends in market stocks. ARMA (p, q) model also used to predict the behavior of KSE-100 index. In this case,
regard, the data under consideration is that of daily closing of KSE-100 cycles from 2015 to 2020 (Cycle 1st to 12th). Descriptive studies of each KSE-100 index cycle is described. Finding the most appropriate models is depending on the least value of Akaike information criterion (AIC), Bayesian Schwarz information criterion (SIC) and Hannan Quinn information criterion (HIC). Durbin-Watson Test also apply for knowing most appropriate model. Root mean square error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error

![Fig. 18 Sample of News Data Sentiment Polarities for the Scrapped News (2015–2020)](image)

![Fig. 19 Predicted Closing index of KSE 100](image)

| Cycles | Duration       | N    | Mean   | Median | Std.D   | Skewness | Kurtosis | Jur-Bera |
|--------|----------------|------|--------|--------|---------|----------|----------|----------|
| 1      | 1.7.2020–31.12.2020 | 128  | 40,498.80 | 40,664.90 | 2055.90 | −0.894793 | 3.56593  | 18.7883  |
| 2      | 1.1.2020–30.6.2020  | 123  | 36,034.35 | 34,181.80 | 4377.675 | 0.155276  | 1.883321 | 6.885004 |
| 3      | 1.7.2019–31.12.2019 | 128  | 34,731.29 | 33,762.00 | 3792.565 | 0.665313  | 2.278830 | 12.21680 |
| 4      | 1.1.2019–28.6.2019  | 122  | 37,646.89 | 38,029.50 | 2224.751 | −0.254873 | 1.939869 | 7.039011 |
| 5      | 2.7.2018–31.12.2018 | 125  | 40,485.38 | 40,869.00 | 1615.475 | −0.467694 | 2.311328 | 7.027196 |
| 6      | 1.1.2018–29.6.2018  | 124  | 43,802.92 | 43,629.00 | 1354.415 | 0.165118  | 2.468991 | 2.020303 |
| 7      | 3.7.2017–29.12.2017 | 125  | 42,075.09 | 41,401.00 | 2318.752 | 0.479395  | 2.290700 | 7.408262 |
| 8      | 2.1.2017–30.6.2017  | 124  | 49,194.90 | 49,038.50 | 1419.163 | 0.154101  | 3.823482 | 3.994414 |
| 9      | 4.7.2016–30.12.2016 | 125  | 41,415.64 | 40,371.00 | 2393.466 | 1.243949  | 3.583683 | 34.01209 |
| 10     | 1.1.2016–30.6.2016  | 128  | 33,999.88 | 33,508.00 | 38,777.00 | 0.415485  | 1.927171 | 9.821190 |
| 11     | 1.7.2015–31.12.2015 | 123  | 34,044.59 | 33,944.00 | 1153.430 | 0.312248  | 2.018026 | 6.940631 |
| 12     | 1.1.2015–30.6.2015  | 126  | 33,284.96 | 33,532.00 | 1087.421 | −1.196852 | 4.822168 | 47.51309 |
MAPE) and Theil’s U-Statistics are used to calculate the forecast accuracy of the best-fitted model of KSE-100 index cycles (1st – 12th). Most of cycle’s follows the most appropriate model is ARMA (2, 1) and some cycles shows the best adequate model is ARMA (3, 1).

Table 2  Diagnostic test of ARMA (p, q) for KSE-100 Index (1–12) cycles

| Cycles | ARMA(p, q) | R² | ADJ R² | SE Reg | Log Likelihood | AIC | SIC | HQC | DWS |
|--------|------------|----|--------|--------|----------------|-----|-----|-----|-----|
| 1      | ARMA (2, 1) | 0.9627 | 0.96184 | 401.6022 | −949.2897 | 14.8952 | 14.9843 | 14.9314 | 1.764681 |
| 2      | ARMA (3, 1) | 0.9547 | 0.95436 | 944.3684 | −1018.787 | 16.6307 | 16.7221 | 16.6678 | 1.29463 |
| 3      | ARMA (3, 1) | 0.9663 | 0.96543 | 705.1330 | −1023.965 | 16.062 | 16.1511 | 16.0982 | 1.254008 |
| 4      | ARMA (3, 1) | 0.9392 | 0.9377 | 555.5077 | −945.5077 | 15.5647 | 15.6566 | 15.6021 | 1.413394 |
| 5      | ARMA (3, 1) | 0.8021 | 0.79716 | 727.569 | −1000.627 | 16.0740 | 16.1645 | 16.1108 | 1.298629 |
| 6      | ARMA (2, 1) | 0.9266 | 0.92478 | 371.4704 | −909.2377 | 14.7295 | 14.8206 | 14.7666 | 1.570193 |
| 7      | ARMA (2, 1) | 0.9523 | 0.95109 | 512.7589 | −1023.965 | 15.3712 | 15.4637 | 15.4099 | 1.977866 |
| 8      | ARMA (1, 1) | 0.8871 | 0.88431 | 842.7006 | −956.8226 | 15.2475 | 15.3384 | 15.2844 | 1.91939 |
| 9      | ARMA (2, 1) | 0.8904 | 0.88771 | 802.0628 | −1012.347 | 16.2616 | 16.3521 | 16.2983 | 1.896819 |
| 10     | ARMA (1, 1) | 0.9858 | 0.98541 | 282.6399 | −904.3153 | 14.924 | 14.8155 | 14.2286 | 1.316453 |
| 11     | ARMA (2, 1) | 0.9333 | 0.93162 | 301.6236 | −876.0780 | 14.3102 | 14.4017 | 14.3474 | 1.719842 |
| 12     | ARMA (4, 1) | 0.7596 | 0.75369 | 539.6835 | −970.8627 | 15.4740 | 15.5641 | 15.5106 | 1.613456 |

Table 3  Forecast Evolution for KSE-100 Index (1–12) cycles

| Cycles | ARMA(p, q) | RMSE | MAE | MAPE | Theil’s U-Statistics |
|--------|------------|------|-----|------|----------------------|
| 1      | ARMA (2, 1) | 4553.89 | 4298.00 | 10.4374 | 0.890773 |
| 2      | ARMA (3, 1) | 4454.441 | 3546.87 | 10.86979 | 0.059424 |
| 3      | ARMA (3, 1) | 3516.736 | 3126.943 | 0.044968 | 0.071229 |
| 4      | ARMA (3, 1) | 2148.940 | 1873.149 | 4.978616 | 0.028672 |
| 5      | ARMA (3, 1) | 1628.932 | 1330.290 | 3.336208 | 0.020801 |
| 6      | ARMA (2, 1) | 1863.329 | 1560.176 | 3.502538 | 0.021564 |
| 7      | ARMA (2, 1) | 2041.976 | 1762.580 | 4.285248 | 0.023928 |
| 8      | ARMA (1, 1) | 1456.558 | 1092.721 | 2.208688 | 0.014848 |
| 9      | ARMA (2, 1) | 2305.544 | 1471.920 | 4.145864 | 0.028073 |
| 10     | ARMA (2, 1) | 2072.524 | 1694.529 | 4.889591 | 0.030573 |
| 11     | ARMA (2, 1) | 969.8036 | 796.4294 | 2.354089 | 0.014202 |
| 12     | ARMA (4, 1) | 1116.435 | 882.2405 | 2.688135 | 0.016771 |

(MAPE) and Theil’s U-Statistics are used to calculate the forecast accuracy of the best-fitted model of KSE-100 index cycles (1st – 12th). Most of cycle’s follows the most appropriate model is ARMA (2, 1) and some cycles shows the best adequate model is ARMA (3, 1).

Fig. 20  Forecast Evolution of cycles (2015–2020)
process apart from cycles 8th and cycle 12th. Cycle 8th shows that the best fitted model is ARMA (1, 1) and cycle 12th follows the adequate model is ARMA (4, 1). MAPE values is less as compare to other forecast evaluations like RMSE and MAE. Theil’s U-Statistics also demonstrate for each cycle. The Theil’s U-Statistics values of each cycles lies approaches to zero shows that the cycle is strongly correlated to previous one.

**Funding** The authors did not receive support from any organization for the submitted work.

**References**

1. Adhikari R (2013) An introduction of the time series modeling and forecasting. (Cornell University library) arXiv.org> cs > arXiv : 1302. 6613
2. Anjum S (2020) Impact of market anomalies on stock exchange: a comparative study of KSE and PSX. Future Business Journal 6(1):1–11
3. Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis: forecasting and control. Prentice Hall, Inc, New Jersey
4. BUSINESS RECORDER (n.d.) [Online]. Available: https://www.brecorder.com/markets/stocks/
5. Cheema A (2008) WEB BASEDStock Forecasters. Springer
6. Daily Times (n.d.) [Online]. Available: https://dailytimes.com.pk/business
7. DAWN (n.d.) [Online]. Available: https://www.dawn.com/business
8. Hegazy O, Soliman OS, Salam MA (2013) A machine learning model for stock market prediction. Int J Comput Sci Telecommum 4:12
9. Hipel KW, McLeod AI (1994) Time Series Modelling of Water Resources and Environmental Systems. Elsevier, Amsterdam
10. Islam RS, Hasan N (2016) Artificial Neural Network Approach for Stock Price and Trend Prediction. Conference: Int Conference Adv Information Commun Technol Chittagong, Bangladesh
11. Islam RS, Sultana N, Hasan N (n.d.) Financial instability analysis using ANN and feature selection technique: application to stock market price prediction. Conference: 2016 international conference on innovations in science, engineering and technology (ICISET)
12. Jain R (2019) Web Scraping a Site with Pagination using BeautifulSoup. [Online]. Available: https://medium.com/analytics-vidhya/webscraping-a-site-with-pagination-using-beautifulsoup-f99a0980445
13. Jenkins JM, Milholland RD, Lilly JP, Beute MK (1970) Commercial gladiolus production in North Carolina. NC Agric Ext Circ 44:1–34
14. Karachi 100 Historical Data (n.d.) [Online]. Available: https://www.investing.com/indices/karachi-100-historical-data
15. Koenig R (2019) NLP for Beginners: Cleaning & Preprocessing Text Data. [Online]. Available: https://towardsdatascience.com/nlp-for-beginners-cleaning-preprocessing-text-data-ae8e30beff0f
16. Moghaddam AH, Moghaddam MH, Esfandyari M (2016) Stock market index prediction using artificial neural network. J Econ
17. Pakistan OBSERVER (n.d.) [Online]. Available: http://pakobserver.net/category/business
18. Rasel RI, Sultana N, Meesad P (2015) An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators. Int J Comput Intell Studies
19. The Express Tribune (n.d.) [Online]. Available: https://tribune.com.pk/kse-100-index
20. Usman M (2016) Stock market prediction using machine learning techniques. In: 2016 3rd International Conference on Computer and Information Sciences (ICCOINS)
21. Yetis Y, Kaplan H, Jamshidi M (2014) Stock market prediction by using artificial neural network. World Automation Congress
22. Yule GU (1927) On the method of investigating periodicities in disturbed series with special reference to Wolfer’s Sunspot numbers. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers Math Phys Char 226: 267–298 (32 pages) published By: Royal Society

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.