Stock Price Prediction Using Machine Learning Methods: An Example of Turkish Banks

MSc Research Project
Data Analytics

Hilal Ozcelik
Student ID: x23218274

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Hilal Ozcelik……………………………………………………………………………………………

Student ID: x23218274……………………………………………………………………………………………..……

Programme: Data Analytics…………………………………………………… **Year:** 2024……………………..

Module: MSc Research Project………………………………………………………………………….………

Supervisor: Jorge Basilio……………………………………………………………………………………….………

Submission Due Date: 12/12/2024……………………………………………………………………………………….………

Project Title: Stock Price Prediction Using Machine Learning Methods: An Example of Turkish Banks………………………………………………………………………………….………

Word Count: 7069……………………………………… **Page Count** …19………………………………………….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Hilal Ozcelik……………………………………………………………………………………………

Date: 12/12/2024………………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Stock Price Prediction Using Machine Learning Methods: An Example of Turkish Banks

Hilal Ozcelik
x23218274

Abstract

This study aims to estimate the stock prices of five major banks in Turkey—Akbank, Garanti, Halk, Is, and Yapı Kredi—that are listed on the Istanbul Stock Exchange. To achieve this, a comprehensive dataset was compiled, including ten years of historical stock prices for these banks, ten years of oil prices, S&P 500 prices, iShares prices, USD/TRY exchange rates, as well as textual data comprising daily notification reports from the respective banks. Utilizing the Long Short-Term Memory (LSTM) method, predictive models were developed under two scenarios: one incorporating all variables, including the textual data, and the other excluding the textual data. The results demonstrate that the LSTM model successfully generated accurate predictions and that the inclusion of textual data positively contributed to the performance of the prediction models.

1 Introduction

Today, the platforms where stocks are bought, sold. and traded are called stock markets. These bought and sold assets can be stocks, bonds, and similar commodities, namely gold, oil or agricultural products. Thanks to these transactions, like buying and selling stocks, the stock market provides capital resources to individual investors and companies while also having a number of economic indicators that meet the fund requirements. In addition to these features, the stock market is based on existing shares and investors' expectations about companies, and these expectations can be affected by many local and global factors. For this reason, stocks on the stock market have a variable and dynamic structure that may have any increasing or decreasing tendency. Although this variable structure makes it difficult to make predictions, it makes the stock market itself attractive for individual investors and companies. With this attraction, the interest in the stock market is increasing, and the research and studies conducted in this field are increasing day by day in parallel.

Stock prices and forecasting have been areas of interest and difficulty for many individual investors, companies, and researchers from the past to the present. Especially in recent years, the size of the data generated as a result of the increase in internet use has increased, so it has become more difficult for investors, companies and researchers to decide which pattern the stocks follow and what the most appropriate decision will be for the stock in question(Ravikumar & Saraf, 2020). In addition, it is said that stock prices have become more complex and variable, affected by local and global events and similar news. In this case, it makes it even more difficult to predict. In this context, trying to make a prediction based only on historical stock prices may not be sufficient(Singh et al., 2019).

Before technology developed as much as it does today and before fields such as machine learning existed, researchers and enthusiasts interested in this field tried to predict stock prices using a number of statistical and econometric methods. However, these methods, which provide successful results in linear problems, have not been as successful as desired in predicting stock prices(Yu & Yan, 2020). In the following years, with the advancement of technology, classical methods with time series such as Linear Regression, ARMA, ARIMA began to be preferred in stock price prediction. Later, machine learning models such as Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), K-nearest Neighbour (KNN) and Naive Bayes (NB) were used to learn and interpret methods. It has started to be used frequently due to its power. Subsequently, in addition to these methods, techniques such as Artificial Neural Network (ANN) and Long-Short Term Memory (LSTM) and Convolutional Neural Network (CNN), which come with deep learning, play an important role among the methods used in predicting stock prices from start to finish(Long et al., 2019). In this study, the prices of the stocks of banks in Türkiye are predicted by using machine learning models mentioned above, and this study seeks an answer to the following research question. And with this research question, the following objectives are implemented respectively.

Research Question: To what extent are the stocks of Turkish banks accurately predicted with the use of LSTM machine learning models? When models are built with textual and non-textual data, in which case are more successful predictions made and which models are better in terms of model performance? Which model and models can be recommended when the results are compared?

Obj 1: Examining the studies that include and do not include sentiment analysis in the last five years, considering the methods they use.

Obj 2: Examination of Türkiye-origin studies containing stock price predictions for all years.

Obj 3: Aggregation of data including stock prices, financial data and text.

Obj 4: Regulating stock prices by taking closing prices and examining variables. Additionally, feature extraction using closing prices.

Obj 5: Implementation of LSTM models for historical stock prices and financial data and evaluation of model results.

Obj 6: Implementation of LSTM models for historical stock prices, financial data and text-containing data.

Obj 7: Comparison of model results with and without text in terms of prediction performance.

In this study, it is of great importance to seek answers to these questions and to realize the stated objectives. In some of the studies in the literature, predictions were made only with historical data on stock prices, and text-based data was not included. In this study, the inclusion of historical stock data, financial indicators and text data makes the study unique, and it also has a unique feature in terms of the Turkish stocks it deals with. Additionally, it is thought that the study is different in that it makes predictions with text-based and non-text-based data and compares the performance of the two situations. Due to these aspects, it is aimed that the study will make a new contribution to the literature and also provide information to investors.

After the introduction section of this study, the studies analysed as a result of the literature research in the second section, the methodology of the study in the third section, then the design in the fourth section, the implementation of the study in line with the methodology and methods specified in the fifth section, the results and evaluations of the implementation in the sixth section, and the conclusion and future research information that can be done in the following years in the last section.

2 Related Work

The literature review shows that many studies have been done on predicting stock prices. The methods and datasets used by these studies may include only historical stock data, or they may contain both historical stock data and text-based data such as social media news or other financial news data, and they differ from each other in this respect. Here, studies conducted in the last five years, between 2019 and 2024, are examined via Google Scholar, and studies conducted in all years for the geographical region of Türkiye have been searched. First, studies that do not contain textual data will be discussed, then studies that contain text will be discussed, and finally studies that predict stock prices in Türkiye will be evaluated.

2.1 Investigating Non-Sentiment Content in Literature

Studies in the literature generally show that they make price predictions or classification predictions as to whether the target variable will increase or decrease by using variables such as opening, closing, highest, lowest and adjusted closing prices. In their study, Naik N. and Mohan Biju R. predicted the prices of stocks of banks such as ICICI Bank, SBI Bank, Yes Bank and Kotak Bank on the Indian stock exchange. In the first stage of their study, the researchers determined the important ones among the 33 variables using the Boruta Algorithm. In the second stage, they developed a solution to this problem by choosing Artificial Neural Network (ANN), Support Vector Machines (SVM) and deep learning algorithms to determine whether the banks' stocks will increase or decrease. As a result, researchers who tackled a classification problem based only on historical stock prices stated that deep learning algorithms performed more successfully(Naik & Mohan, 2019). In another study, Nikou et al. used only historical stock prices. They preferred to use stocks on the UK stock exchange in their studies. Researchers made predictions for the future by taking the daily closing prices of the stocks they determined between January 2015 and June 2018, using the Artificial Neural Network (ANN), Random Forest (RF), LSTM (Long-Short Term Memory) and Support Vector Machines (SVM) methods. They wanted to determine the method that makes the prediction. They compared the results of the models according to the evaluation criteria of Root Meen Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE) and found that the LSTM method had less error in predicting the prices of the stocks they determined. In addition to the results of their study, researchers also stated that more successful results can be obtained in the future with different types of LSTM model, such as bi-directional LSTM or encoder-decoder LSTM(Nikou et al., 2019). In their stock price direction prediction study, Zhong and Enke classified the daily price direction of the asset coded SPY on the New York Stock Exchange. For this purpose, the authors carried out their prediction studies in two phases and cleared the data such as missing values and outliers for ten years of historical data. Then, they built the first prediction models with 60 variables using ANN and DNN algorithms for these clean data sets. As a second stage, they applied the PCA method to the clean datasets, applied the variable reduction method, and built their models in this way. As a result, the DNN model trained using PCA performed slightly better than the other model(Zhong & Enke, 2019). Yu P. and Yan. in their studies in 2020, they made predictions for S&P500, DJIA, Nikkei 225, HSI, CSI300 and ChiNext stocks. Researchers used the time series-based PCR method along with deep neural networks (DNN) and LSTM in their studies. In addition, researchers using time series methods such as ARIMA, SVM, Multi-Layer Perceptron (MLP) compared the model outputs in their studies, taking into account criteria such as accuracy, RMSE, mean absolute error percentage (MAPE), Correlation coefficient (Corr). As a result, they found that the combination of LSTM and PCR methods exhibited the best prediction performance(Yu & Yan, 2020). Another researcher Moghar A. and Hamiche M., who used only historical stocks in their studies, tried to predict the opening prices of NKE and GOOGL assets on the New York Stock Exchange from historical stock data. Researchers used Recurrent Neural Network (RNN) and LSTM deep learning models in prediction studies. Moghar A. and Hamiche M., who run these models at different epoch numbers, stated that as the number of epochs increases, the error rates of the models decrease and their prediction abilities increase(Moghar & Hamiche, 2020). In another study, Nabipour et. al. used Decision Tree, Bagging, Random Forest, Adaboost, Gradient Boosting, XGBoost, ANN, RNN and LSTM models, they grouped the assets in the Iranian Stock Exchange on a sectoral basis and made a sector-based stock price prediction. Researchers, who classified stocks in four groups as finance, oil, non-metallic minerals and base metals, made predictions based on 10-year historical stock prices for 1, 2, 5, 10, 15, 20 and 30-day periods. The authors, who compared the prediction results according to evaluation criteria such as MAPE, MAE, MSE and RMSE, stated that the LSTM model gave the best results among all methods(Nabipour, Nayyeri, Jabani, Mosavi, et al., 2020). In the same year, Nabipour et al. in another study again used historical data on stocks on the Iranian stock exchange, but made a two-phase prediction. In this study, which took place in two different phases, the authors preferred Decision Tree, Random Forest, Adaboost, Naïve While Bayes, K-Nearest Neighbors (KNN), Logistic Regression, ANN, XGBoost, Support Vector Classifier (SVC), RNN and LSTM methods. In the first phase, the authors predicted historical stock prices as continuous data, and in the second phase, they treated the target variable as binary and built their models. Comparing the results according to metrics such as accuracy, F1-score and ROC-AUC, the authors stated that RNN and LSTM techniques were more successful in both phases, and found that the prediction performance in continuous data was slightly better than binary prediction results(Nabipour, Nayyeri, Jabani, Shahab, et al., 2020). In another study, Shen J. and Shafiq O. made their study unique by collecting extensive data and adding not only historical stock prices but also a number of different data such as business and stakeholder data. In their study, the authors predicted the prices of assets in the Chinese Stock Exchange and also used the variable selection and Principal Component Analysis (PCA) method to predict the prices of the relevant assets with the LSTM model. Comparing their results with the results of previous studies, the authors stated that the prediction performance of their models was quite good(Shen & Shafiq, 2020). In their study, Yadav et al. used four assets on the Indian Stock Exchange and predicted their prices. The authors who used the LSTM model in their studies stated that there were no parameters determined for the LSTM model and that they found the most appropriate parameters by model tuning. In this regard, the authors who used stateful and stateless LSTM models for four different stocks in their first experiments stated that stateless LSTM was more successful, and in the second experiment, the authors who tried different hidden layers from one to seven stated that they got the best result when this value was 1(Yadav et al., 2020).Sunny et al. by taking historical data of the GOOGL asset on the New York Stock Exchange, they predicted the future prices of the asset using RNN, LSTM and bi-directional LSTM methods. Researchers tried all these models for different hidden layers and epoch numbers and determined the features that showed the best performance and evaluated the results of the models by looking at the RMSE error results. According to their evaluations, the bi-directional LSTM method has fewer errors and has been emphasized as a method that researchers and investors should prefer(Istiake Sunny et al., 2020). Mukherjee et al. pointed out the difficulty of predicting stock prices in their studies and emphasized that there are many studies in this field and deep learning models such as ANN and CNN are used for prediction. In their study, unlike other studies, the authors obtained historical stock data for the assets they determined from the India National Stock exchange between 2008 and 2018, and in addition, they added recent data between 2019 and 2020 to their dataset. Stating that their study is different from previous ones in this aspect, the authors performed the prediction process using ANN and CNN algorithms and stated that the CNN model performed better with 98.92% accuracy(Mukherjee et al., 2023). In their study, Han and Fu pointed out the difficulty of stock price prediction and made predictions about the future based on ten-year historical data of Appl stock. Researchers who preferred the Bi-LSTM algorithm for prediction applied min-max scale before training the data set and then trained their models. The researchers, who showed the prediction results on a visual in the form of actual and predicted in their studies, also stated that the Bi-LSTM model is an effective method in terms of capturing trends and patterns in the dataset and can be used for future predictions(Han & Fu, 2023). Sen et al. conducted a study on determining the direction of stocks in the Indian National Stock Exchange. According to the thesis defended, although stock prices are said to be unpredictable, the authors argue that prediction is possible with a high accuracy rate. While using CNN and LSTM algorithms in their studies, they took data between 2014 and 2020 and used the multi-step prediction method and walk-forward validation method. As a result of their studies, they stated that although CNN runs faster, LSTM exhibits superior performance(Sen et al., 2023).

2.2 Investigating Sentiment Content in Literature

In the previous section, studies that made predictions using only historical stock data were examined, and it was stated in these studies that adding text content and similar data to the study in addition to historical stock data would increase the forecast performance. In this section, studies that use text data in addition to historical stock data are examined. Mokhtari et al. in their study, they drew attention to the importance of using textual data and used the textual data they received from Twitter in the problem of classifying stocks. In addition to predicting the price of stocks, they also conducted a three-class forecast study on whether to buy, sell or hold shares. Models such as Bernoulli Naive Bayes (BNB), Decision Tree (DT), Logistic regression (LR), Gaussian Naive Bayes (GNB), Random Forest (RF), KNN, SVM, XGBoost and ANN are used for the classification problem, and LSTM and ANN are used for price prediction. while Linear Regression was more successful in terms of RMSE, MAPE, MSE, R^2 evaluation metrics in stock price prediction, they determined that the SVM model made a better prediction with 76% accuracy for the classification problem(Mokhtari et al., 2021a). Sirimevan et al. like the previous study, they emphasized the addition of textual news, events, and social media comments to forecasting studies. For their study, in addition to the historical data of APPL stock, they collected data between 2016 and 2019 on Twitter and other textual data from other search engines. The authors, who preferred RNN-LSTM models and made predictions for each data and data group, compared all daily prediction results and shared their results. They stated that model performances could yield more successful results by adding other variables such as gold and foreign currency in their studies in the following years(Sirimevan et al., 2019). In another study, Jin et al. predicted the closing prices of APPL stock. For this purpose, in addition to historical stock prices of APPL stock, they also used users' comments about the stock and also added LSTM methods, which they combined with CNN, LSTM and other methods, to their studies. In the first stage, the authors made predictions using the CNN method based on user comments. In the second step, since stock data includes the time component, they applied empirical modal decomposition (EMD) using historical stock prices. In the last step, they applied LSTM and revised LSTM methods to the datasets. At the end of their study, the authors touched on the advantages and disadvantages of the methods they used and stated that more successful results could be obtained with different methods that can be applied in the future(Jin et al., 2020).In another study, Ji et al. They included historical stock prices and social media news in their studies. The authors combined textual and financial data using Doc2Vec and Stacked Autoencoder (SAE) to process text data and carried out a two-phase study with the LSTM model. They implemented the first phase of their study with only financial data and the second phase with financial plus text content data. Comparing their results with MAE, RMSE and R2, researchers stated that using text data in addition to numerical data reduced error rates and showed better prediction performance(Ji et al., 2021). Finally, Khan et al. in their studies, they did not limit themselves only to historical stock prices and also included financial news data and social media news in their studies. Using the stocks of companies such as Oracle, Nokia and Microsoft in their studies, the authors tried to solve a classification problem as to whether the direction of the stocks would be positive, negative or neutral. For this classification problem, the authors used GNB, Multinomial Bayes (MNB), SVM, LR, MLP KNN, Classification and Regression Tree (CART), Linear Discriminant Analysis (LDA), AdaBoost, Gradient Boosting Classifier (GBM), RF and Extra Tree (ET)) models and compared the model results by looking at evaluation criteria such as precision, recall and F-score. As a result, they found that the Random Forest (RF) model gave superior results compared to other models(Khan et al., 2022). In another study, Shilpa and Shambhavi added historical stock prices as well as stock-related news texts to their study, making their study unique in terms of the methods they used. In their study, they extracted financial indicators such as moving average convergence divergence (MACD), relative strength index (RSI) and moving average (MA) from historical stock prices and added them to their variables. Then, in order to convert the news texts into numerical data, they pre-processed the data, performed operations such as stemming and tokenization, and also performed feature extraction. In their study, the authors applied this data to six different data sets on a daily, monthly and annual basis while making predictions for the assets of two different companies. As a result, the authors, who suggested a different model in their study, stated that the NN + DBN + SIWOA (Neural Network + Deep Belief Network + Self-improved whale optimization algorithm) combined model is superior(Shilpa & Shambhavi, 2023). In another study, Zhao and Yang predicted the direction of shares of the HSI asset on the Hong Kong Stock Exchange. The authors, who proposed a model called SA-DLSTM for this prediction, used emotion enhanced convolutional neural network (ECNN), the denoising autoencoder (DAE) models, and LSTM models. In their study, the authors, who added the comments of users on the internet to their data sets in addition to historical stock prices, increased the accuracy of the model by performing feature extraction and stated that the model they proposed would be more useful for investors and others in decision-making(Zhao & Yang, 2023).

2.3 Uncovering Turkish Sources

In this section, the studies examined in the literature review based on the geographical region of Türkiye are mentioned. Emir et. al. in their study, they predicted the stock prices of eight banks in the Istanbul Stock Exchange and used a dataset including technical variables for prediction. While determining the variable that contributed the most to the prediction, they also established prediction models with Random Forest Regression and ANN methods. The authors compared the results of their models considering MAE, MSE and Median Absolute Error (MedAE) and found that the ANN method showed the best performance(Emir et al., 2016). In another study, Gunduz determined the direction of the stocks by using the hourly historical stock prices of eight banks on the Istanbul Stock Exchange. Carrying out a two-phase study, Gunduz used all variables in the first stage, while retraining the variable-reduced models in the second stage. Gunduz, who used SVM, Light GBM and LSTM methods in his study, stated that he achieved more successful results with the LSTM algorithm and variable reduction(Gunduz, 2021). Another study is the stock price prediction study of three banks on the Istanbul Stock Exchange in 2021 by Eguz et al. In this study, Eguz et al. predicted the prices of these stocks with a total of 36 variables, including historical stock data and other variables of Akbank, Garanti and Yapi Kredi banks between 2016 and 2021. In prediction studies, researchers who preferred multiple linear regression, ridge regression, lasso regression, support vector machines, decision tree models, random forest and XGBoost machine learning models compared the results by looking at MSE and MAE results and stated that XGBoost showed the best performance(Egüz et al., 2021). Kocaoglu et. al. made a sector-based evaluation in its studies and first grouped the stocks on the Istanbul Stock Exchange according to sectors such as white goods, iron and steel, petrochemicals, transportation and communication. In their study, the authors carried out a prediction study using XGBoost, Support Vector Machines (SVM), K-nearest neighbors (KNN) and Random Forest (RF) algorithms and reported the model results. They compared them in terms of criteria such as R-Squared, MAPE and mean squared error (MSE). The authors stated that as a result of these steps, they made the best predictions in the fields of iron and steel and petrochemicals(Kocaoglu et al., 2022).Finally, in another study, Ilkucar predicted the prices of Turkish Airlines shares using FNN, LSTM and GRU methods. The author, who used stocks between 2010 and 2022 in his study, compared the results of the models by looking at R-squared, MSE, RMSE and MAE values, and according to the R-square devaluation metric, he achieved 97% success in the FNN model and 99% success in LSTM and GRU (Ilkucar, 2023).

As a result of the literature review, twenty-five different studies were examined for this study. In these studies, it has been observed that SVM, DT, RF and similar machine learning models and a combination of ANN, DNN, CNN, LSTM and similar models are used when predicting stock prices. While researchers often use historical stock prices for predictions, it has been found that text-based data is also considered as a data set in some studies. It is thought that there is a gap in the literature in this area where text content data is not used in studies conducted in the geographical region of Türkiye, and the study will fill this gap in this aspect. At the same time, it is anticipated that this study will both contribute to the literature and be a source of information for investors in the relevant stock market.

3 Research Methodology

It is seen that two different approaches are preferred in research conducted in computer science. These approaches are listed as CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, and KDD, which stands for Knowledge Discovery in Data. In this research, the KDD approach is used because it does not have a business step. In this third part of the research, the data sets used in the research, how and where these data sets were obtained, the methodology of the research and the project design process flow are mentioned, respectively.

3.1 Data Resources

The data sets included in this project are collected from three different websites. Firstly, historical stock data of five different Turkish banks called Akbank, Garanti, Halkbank, Isbank, Yapi Kredi were obtained from the Yahoo Finance website, including opening, closing, lowest and highest prices.[1] These datasets cover data between September 1, 2014 and August 30, 2024 and constitute dataset 1. The second dataset includes financial indicators and was taken from a website of Polish origin.[2] The content of the second dataset taken from this website includes USD/TRY exchange

[1] Yahoo Finance https://finance.yahoo.com/

[2] https://stooq.com/

rates, daily crude oil prices, iShares MSCI Turkey ETF (TUR.US) data, that is, measurement values that track the performance of companies traded on a Turkish Stock Exchange, and S&P 500 - U.S. (^SPX) generates daily data tracking the performance of US 500-listed companies. Finally, the third data group included in the research is the data containing daily notification data of each Turkish bank and was obtained from a website originating from Türkiye.[3]

3.2 Methodology

The data sets included in the research are collected from three different sources, and which data sets were taken from which sources were mentioned in the previous section. After these three data sets are obtained, machine learning models are applied, as seen in Figure 1.

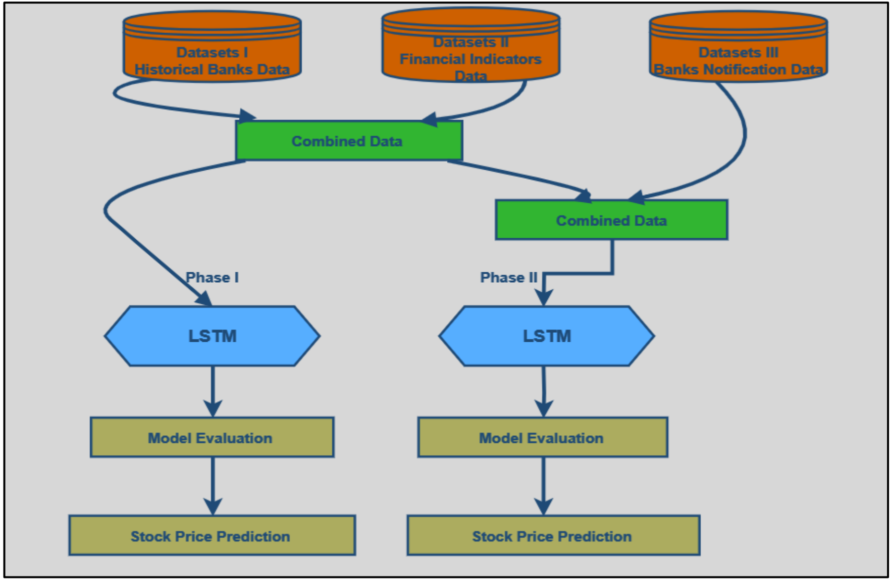


Figure 1: Methodology of Stock Price Prediction.

For this purpose, in the first phase, historical stock data, that is, the first group dataset and the second group dataset containing financial indicators are brought together and LSTM models are applied. In the second stage, all data groups are brought together and the same machine learning models are applied again. The machine learning models to be used in this study were selected by the researcher. Detailed information about the models used is provided below.

LSTM, one of the models to be applied in this study, in other words Long-Short Term Memory, is a different version of the RNN model. To understand LSTM, it is first useful to know the RNN algorithm. RNN is an algorithm from the deep learning family, it has a structure with feedback loops. However, since RNN has vanishing and exploding problems, it may stop learning or require a high learning coefficient. Therefore, an alternative such as LSTM prevents such problems and provides successful results in text content and time series(Sen et al., 2023).

LSTM typically consists of three structures. These are input gate, forget gate and output gate(Sonkavde et al., 2023). While all these three gates use the sigmoid activation function, the input gate decides which data should be entered, the forget gate decides which data should be deleted and the output gate decides which data should be output(Yadav et al., 2020). The LSTM model is frequently used in stock price prediction studies due to its architecture and the fact that it avoids the problems caused by RNN (Istiake et al., 2020; Jin et al., 2020; Moghar & Hamiche, 2020; Sarode et al., 2019; Yadav et al., 2020).Considering all these advantages, it is inevitable for this study that stocks have time content and that the LSTM model, which makes good predictions for the near future, is used.

[3] https://www.kap.org.tr/en/

Once the models are created and trained, all models need to be evaluated. In the literature, it is seen that researchers use different evaluation metrics such as MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), ROC curve(Lu et al., 2021; Mokhtari et al., 2018). In this study, models are evaluated by looking at RMSE (Root Mean Square Error) and R^2 (R-Square) evaluation metrics. RMSE is obtained by dividing the square of the difference between predicted values and actual values by the sample size(Mokhtari et al., 2018). And the RMSE value is calculated through the equation below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

R-Squared is defined as a statistical evaluation metric that shows how much of the variability in the dependent variable is explained by the independent variable in the model(Mokhtari et al., 2018). And it can be calculated with the help of the equation below.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

In this study, the model evaluation metrics explained above with their equations are used in the implementation part of the study. Additionally, after all these steps, predictions are made with the finally established machine learning models.

4 Design Specification

The project basically consists of three stages. As seen in Figure 2, these are data generation, that is, first collecting ten-year data on each bank's stocks and creating the first group data set. In the second step, USD and Turkish Lira exchange rates, crude oil prices, iShares MSCI Turkey ETF indicators and S&P 500- U.S. (^SPX) data are collected and these constitute the second group dataset. Finally, it can be said that daily notifications about each bank is taken from the relevant source and these constitute the third group dataset. In the data engineering phase, which is the second phase of the project, data pre-processing stages such as cleaning of missing data, normalization and feature extraction are carried out, and some visualizations are also used to better understand the data. In this section, which consists of two phases, the merging of the first group and second group data represents the first phase, while the second phase refers to the merging of all data.

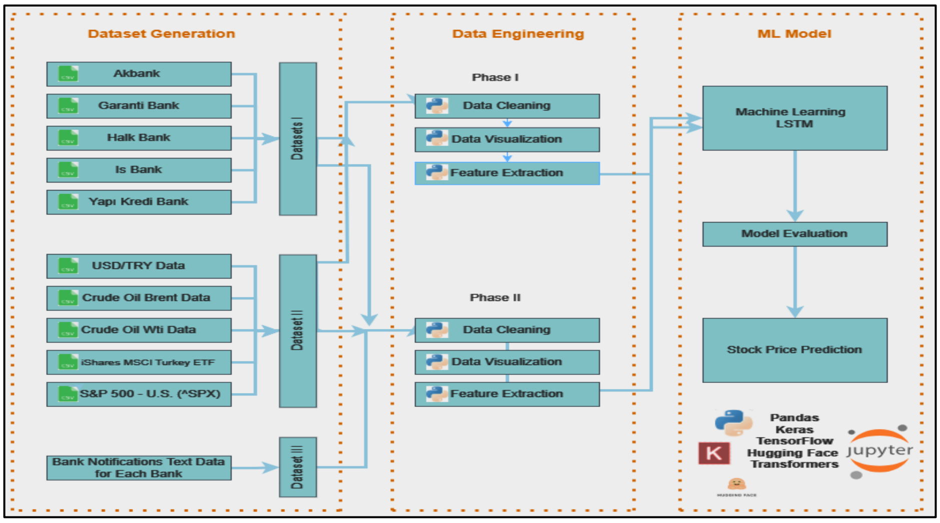


Figure 2: Project Design Framework to Stock Price Prediction.

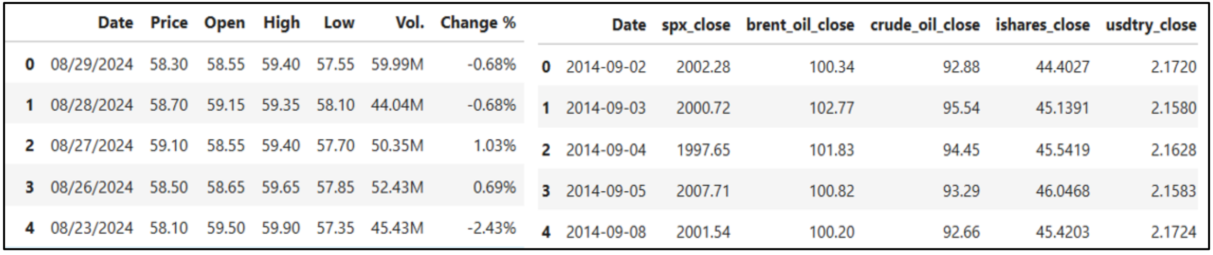
Additionally, the transformers library provided by Hugging Face is used to convert textual notifications about banks into numerical data. After all these steps, the third and final phase of the project comes to the establishment of machine learning models and evaluation of the results. In this step, the same LSTM models are established for both the first and phases stages and the model outputs are evaluated with evaluation criterias.

5 Implementation of Stock Price Prediction for Turkish Banks

In this section, what was done in the implementation part of the study is explained step by step, taking into account the previously explained methodology and design stages. In this regard, this part of the study is explained under three headings: preparation of datasets and pre-processing, feature extraction and obtaining final datasets, sentiment analysis and modelling.

5.1 Preparation and Pre-Processing of Each Datasets

During the preparation of the datasets, historical stock prices data were uploaded separately in csv format for Akbank, Garanti, Halk, Is and Yapi Bank. As seen in Figure 3 in Akbank's dataset, all historical stock datasets contain date, price, that is, closing, opening, high, low, vol. and change % have data containing the variables. Additionally, these datasets cover data between September 1, 2014 and August 30, 2024.



	Date	Price	Open	High	Low	Vol.	Change %
0	08/29/2024	58.30	58.55	59.40	57.55	59.99M	-0.68%
1	08/28/2024	58.70	59.15	59.35	58.10	44.04M	-0.68%
2	08/27/2024	59.10	58.55	59.40	57.70	50.35M	1.03%
3	08/26/2024	58.50	58.65	59.65	57.85	52.43M	0.69%
4	08/23/2024	58.10	59.50	59.90	57.35	45.43M	-2.43%

	Date	spx_close	brent_oil_close	crude_oil_close	ishares_close	usdtry_close
0	2014-09-02	2002.28	100.34	92.88	44.4027	2.1720
1	2014-09-03	2000.72	102.77	95.54	45.1391	2.1580
2	2014-09-04	1997.65	101.83	94.45	45.5419	2.1628
3	2014-09-05	2007.71	100.82	93.29	46.0468	2.1583
4	2014-09-08	2001.54	100.20	92.66	45.4203	2.1724

Figure 3: Akbank Historical Stock Prices Dataset and Financial Dataset.

Similarly, for the same date ranges, USD/TRY, brent oil, crude oil, iShares MSCI Turkey ETF (TUR.US), and S&P 500 - U.S. (^SPX) closing prices are combined into a single dataset and called financial dataset. This dataset can also be seen in Figure 3. The dataset containing historical stock data and financial data for each bank was combined and made ready for the next stage. In addition to gathering and preparing datasets, it is useful to use some visuals to better understand the data and see patterns in the data. For this purpose, relevant information about how banks' closing prices have changed over time was obtained from Figure 4.

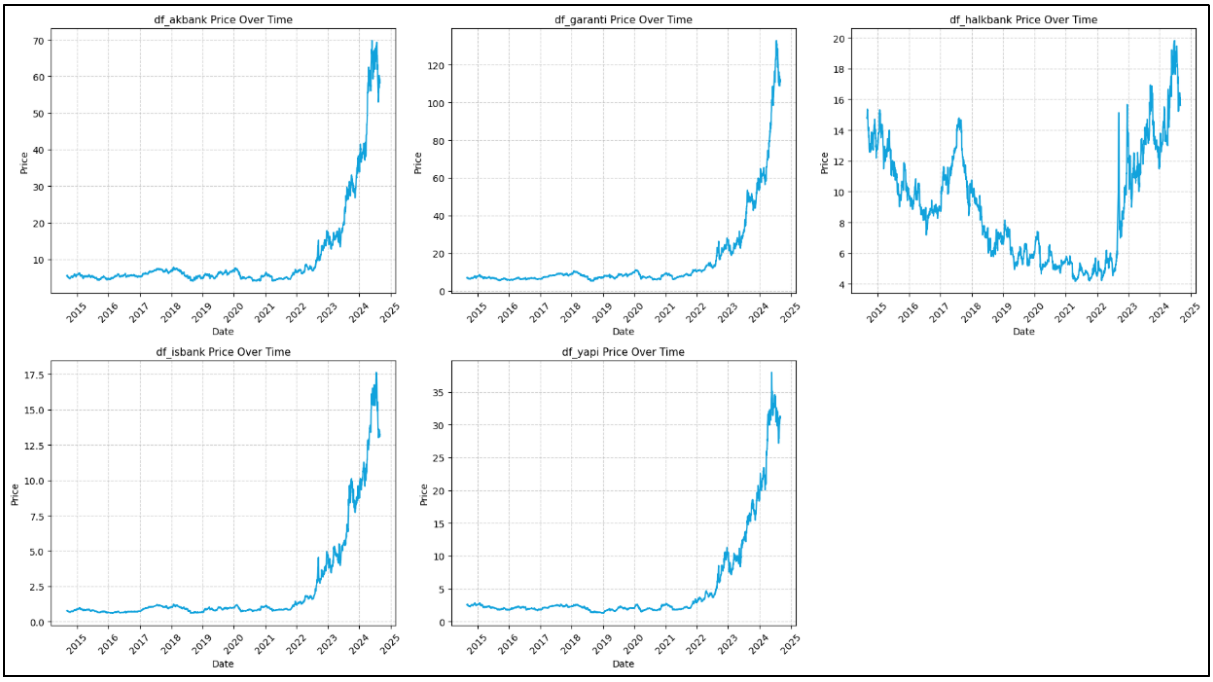


Figure 4: Historical Bank Closing Prices over Time.

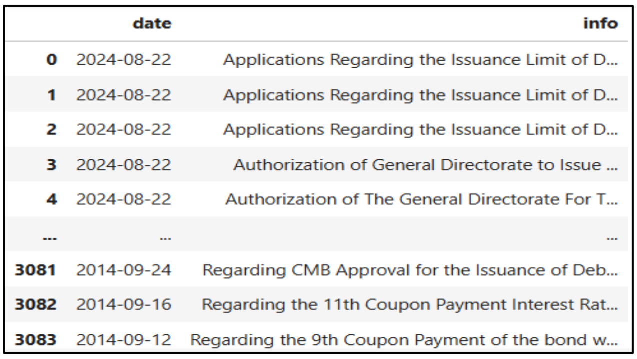
For the pre-processing phase of the data, first the data types were examined, then it was checked whether there were missing values or not, and the missing values were removed because they constituted a small part of the data. Additionally, it was checked for duplicate values and the data was converted to the appropriate format.

5.2 Feature Extraction: Generating the Final Datasets

Feature extraction covers all operations performed by obtaining new variables from existing variables. For this purpose, studies in the literature that extract features from stock price datasets were examined. It has been observed that features such as MA, that is, Moving Averages, SMA (Simple Moving Average), EMA (Exponential Moving Average), RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence), MACD signal, MACD histogram, have been extracted(Jiang, 2021; Mokhtari et al., 2021b; Shilpa & Shambhavi, 2023). The extraction of these features was carried out in the feature engineering section. In addition, features such as average price, price difference, lag features and date, month and year were extracted and performed separately for each dataset. As a result of all these feature extractions, at this stage, the datasets of all banks consist of 10-year daily historical stock prices and 41 variables.

5.3 Sentiment Analysis and Modelling

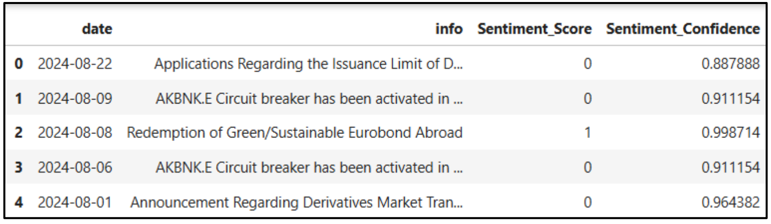
Although historical stock price data and financial variables are used to predict stock prices, studies in the literature have shown that local or global news, political events, users' comments, or other different textual content on social media also used to predict stock prices or the direction of prices. And this offers very successful solutions(Jin et al., 2020; Mokhtari et al., 2021b; Sirimevan et al., 2019). Particularly in the study conducted by Jin et al. in 2020, attention is drawn to the importance of the correlation between user comments and stocks(Jin et al., 2020). Accordingly, in this study, the daily bank notifications of a website originating from Türkiye, which publishes daily bank notifications about the relevant Turkish banks, were queried in English for 10 years. An example dataset of bank notifications is seen in Figure 5.



	date	info
0	2024-08-22	Applications Regarding the Issuance Limit of D...
1	2024-08-22	Applications Regarding the Issuance Limit of D...
2	2024-08-22	Applications Regarding the Issuance Limit of D...
3	2024-08-22	Authorization of General Directorate to Issue ...
4	2024-08-22	Authorization of The General Directorate For T...
...	...	...
3081	2014-09-24	Regarding CMB Approval for the Issuance of Deb...
3082	2014-09-16	Regarding the 11th Coupon Payment Interest Rat...
3083	2014-09-12	Regarding the 9th Coupon Payment of the bond w...

Figure 5: Daily Bank Notifications Dataset Example.

Here, date contains the date when the relevant notification was made, and info contains the notification text. For each bank's notification datasets, rows that did not have information on the specified dates were removed from the dataset and info column for the same dates were merged. Afterwards, sentiment analysis was performed with the DistilBERT model in the Transformers library provided by Hugging Face. Thus, the positive or negative tone of the notifications in English was detected and added to the data set as a sentiment score. In addition, the sentiment score variable is labeled for whether the texts are 1, meaning positive, or 0, meaning negative. The image of an analysed bank notification dataset can be seen in Figure 6.



	date	info	Sentiment_Score	Sentiment_Confidence
0	2024-08-22	Applications Regarding the Issuance Limit of D...	0	0.887888
1	2024-08-09	AKBNK.E Circuit breaker has been activated in ...	0	0.911154
2	2024-08-08	Redemption of Green/Sustainable Eurobond Abroad	1	0.998714
3	2024-08-06	AKBNK.E Circuit breaker has been activated in ...	0	0.911154
4	2024-08-01	Announcement Regarding Derivatives Market Tran...	0	0.964382

Figure 6: Daily Bank Notifications Example with Sentiment Confidence.

After all the datasets were prepared, the datasets of each bank were divided into two: for train data from September 2014 to September 2021, and for test data from September 2021 to September 2024. Then, scaling was performed with MinMaxScaler and the first phase of the implementation was carried out without text. During the modelling phase, the model was applied with the LSTM model. In this model, the time period was specified as 30 days, 50 neurons were preferred, a 0.2 droupout layer was added to prevent overfitting and adam optimizer was used. For the training of the models, Early stopping has been added, which will stop if there is no improvement in 10 consecutive epochs, the batch size is 32, the number of epochs is 50 for Akbank, Garanti and Halk, and 70 for Is and Yapi Kredi banks.

In the second phase of the implementation, the same train-test split process was performed, the same models were established with the same parameters and similar conditions were created. At this stage, sentiment confidence values obtained as a result of sentiment analysis were added to each bank dataset and the applications were carried out in this way.

6 Results and Evaluation of Stock Price Prediction Models

In this section, how prediction models predict bank stocks is shown on graphs and the model evaluation metrics are compared in terms of their results.

6.1 Results and Evaluation of LSTM models for Phase I and Phase II

As it is known, in this study, stock prices of five Turkish banks were predicted in the first phase without including text data, and in the second phase, by including text data. For this purpose, the data was divided into two as the train set from 01.09.2014 to 01.09.2021 and the test set from 01.09.2021 to 01.09.2024, and LSTM models were established under the same conditions for these two stages. In Figure 7, on the left, the results of predicting the stocks of five banks with the LSTM model without text data are shown with the help of a line chart as actual vs predicted price. Similarly, the results obtained as a result of including text data and re-establishing the LSTM models are shown on the right side of Figure 7.

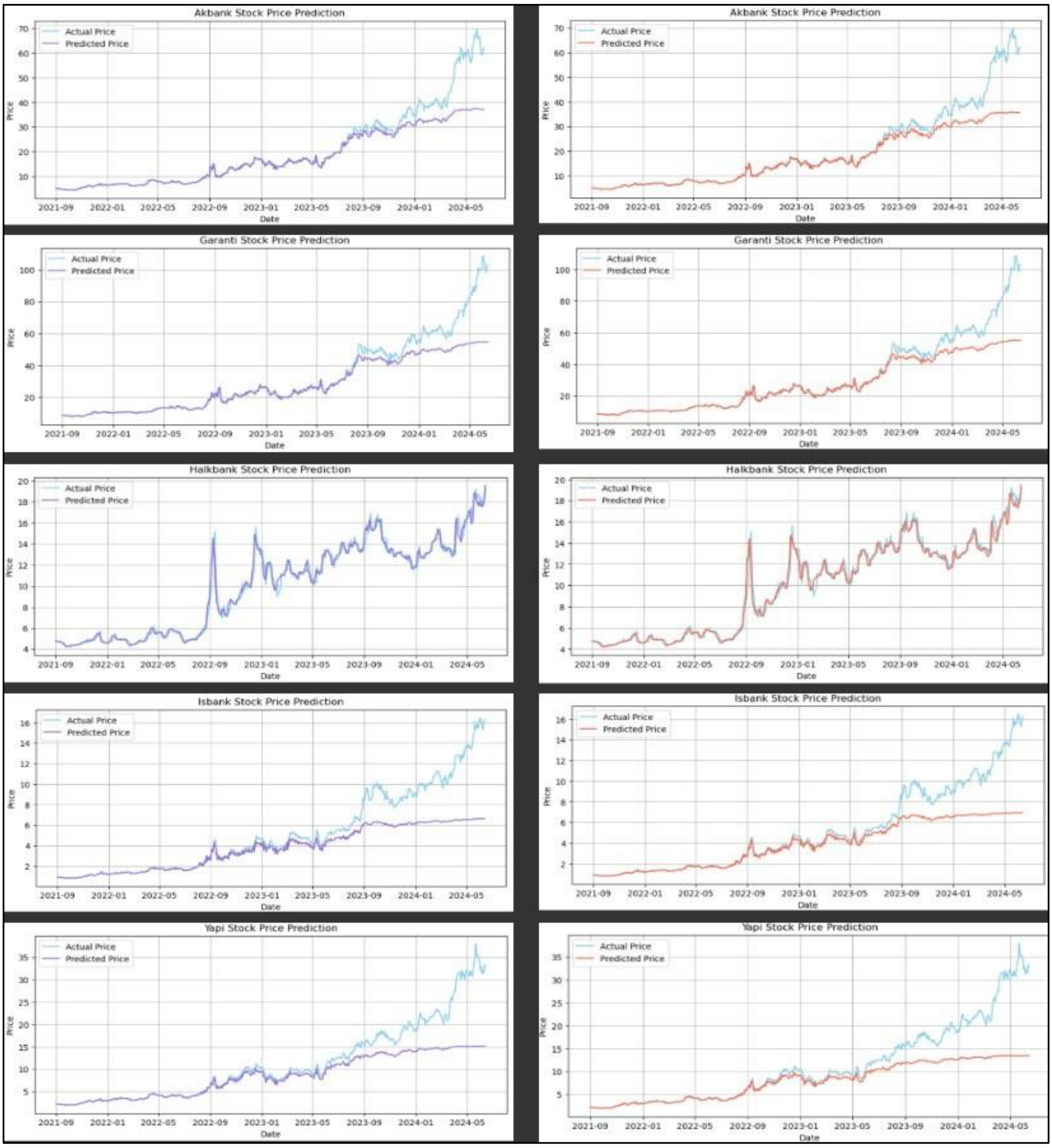


Figure 7: Bank Stocks Price Prediction LSTM Models Actual Vs Predicted Price.

It is possible to easily say that the LSTM model makes the best predictions for Halk by looking at the graph. On the other hand, it is useful to look at the results of the model evaluation metrics in order to better interpret the results. For this purpose, Table 1 shows the test RMSE and R-Squared values of the LSTM models established for each bank in phase one. Similarly, the results obtained as a result of the models built by including the LSTM and text data of these five banks can be seen in the phase two section of the table.

Table 1: Phase I and Phase II LSTM Results

	Phase I		**Phase II**	
Bank	**Test RMSE**	**R-Squared**	**Test RMSE**	**R-Squared**
Akbank	7.736	0.75	7.140	0.79
Garanti	10.881	0.78	10.339	0.80
Halk	0.521	0.98	0.565	0.98
Is	2.909	0.45	2.403	0.63
Yapi	5.516	0.57	6.038	0.48

6.2 Discussion

In this study, a total of ten LSTM models were established for five different Turkish banks, Akbank, Garanti, Halkbank, Isbank and Yapı Kredi, in two phases in total. A forecast study was also conducted by Eguz et. al. in 2021 regarding the stock prices of Akbank, Garanti and Yapı Kredi banks, which are traded on the Istanbul Stock Exchange. However, Eguz et al. did not use any textual data in the study and stated that among the models they used, XGBoost showed the best performance in terms of MAE and MSE metrics(Egüz et al., 2021). In addition to numerical data, text data was also used in this study. And in chapter one, prediction success was observed in cases with and without text, which is one of the research objectives mentioned. According to the model outputs, it can be said that text data contributes positively to the prediction. However, what type of text data and how much text data is used is also important here. Although it is said in the literature that news texts on social media or comments made by users about stocks are used and that they have a positive role in the prediction results, there is no study on the use of bank notification texts as in this study(Jin et al., 2020). Therefore, it would not be appropriate to compare the results of this study with any previous study.

7 Conclusion and Future Work

In this research, stock prices of five Turkish origin banks selected by the researcher were predicted. In this regard, first of all, stock price prediction studies in the literature were examined as studies with or without text data and they gave direction for the realization of this study from beginning to end. Thus, the first of the research objectives was completed. In the literature review, especially studies originating from Türkiye were examined and it was determined that banks' stock prices were predicted, but no prediction was made with text data. In this study, with the LSTM model without text data, the test RMSE value for Akbank is 7.736, the R-Squared value is 0.75, the RMSE value for Garanti is 10.881, the R-Squared value is 0.78, the RMSE for Halk is 0.521, the R-Squared value is 0.98, the RMSE for Is is 2.909, the R-Squared value is 0.75, respectively. RMSE for Squared 0.45 and Yapı Kredi 5.516, R-Squared calculated as 0.57. In the results obtained under the same conditions but with text data, the test RMSE value for Akbank is 7.140, the R-Squared value is 0.79, the RMSE value for Garanti is 10.339, the R-Squared value is 0.80, the RMSE for Halk is 0.565, the R-Squared value is 0.98, the RMSE for Isbank is 2.403, R-Squared 0.63 and RMSE for Yapi Kredi Bank 6.038, R-Squared was obtained as 0.48. According to these results, it can be said that text data has a positive impact on the prediction model.

This study was limited to stocks originating from Türkiye and a comparison could not be made since no similar study had been conducted with the same stocks and methods. However, in the academic field, many studies have been conducted with certain stocks and methods, both with and without text data. In the future work, more successful results can be obtained by doing a similar study in different ways, using different stocks, trying different machine learning models and parameters, adding different variables, or obtaining text data from social media platforms such as Twitter.

References

Egüz, B., Çorbacı, F. E., & Kaya, T. (2021). Stock Price Prediction of Turkish Banks Using Machine Learning Methods. *Lecture Notes in Networks and Systems*, *308*, 222–229.

Emir, S., Dincer, H., Hacioglu, U., & Yuksel, S. (2016). Random Regression Forest Model using Technical Analysis Variables. *International Journal of Finance & Banking Studies (2147-4486)*, *5*(3), 85–102.

Gunduz, H. (2021). An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. *Financial Innovation*, *7*(1).

Han, C., & Fu, X. (2023). Challenge and Opportunity: Deep Learning-Based Stock Price Prediction by Using Bi-directional LSTM Model. *Frontiers in Business, Economics and Management*, *8*(2), 51–54.

Ilkucar, M. (2023). Prediction Turkish Airlines BIST Stock Price Through Deep Artificial Neural Network Considering Transaction Volume and Seasonal Values. *Bilişim Teknolojileri Dergisi*, *16*(1), 43–53.

Istiake Sunny, M. A., Maswood, M. M. S., & Alharbi, A. G. (2020). Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. *2nd Novel Intelligent and Leading Emerging Sciences Conference, NILES 2020*, 87–92.

Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. *International Journal of Crowd Science*, *5*(1), 55–72.

Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. In *Expert Systems with Applications* (Vol. 184). Elsevier Ltd.

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, *32*(13), 9713–9729.

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, *13*(7), 3433–3456.

Kocaoglu, D., Turgut, K., & Konyar, M. Z. (2022). Sector-Based Stock Price Prediction with Machine Learning Models. *Sakarya University Journal Of Computer And Information Sciences*, *5*(3).

Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, *164*, 163–173.

Lu, W., Li, J., Wang, J., & Qin, L. (2021). A CNN-BiLSTM-AM method for stock price prediction. In *Neural Computing and Applications* (Vol. 33, Issue 10, pp. 4741–4753). Springer Science and Business Media Deutschland GmbH.

Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. *Procedia Computer Science*, *170*, 1168–1173.

Mokhtari, S., Yen, K. K., & Liu, J. (2018). *Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning*.

Mokhtari, S., Yen, K. K., & Liu, J. (2021a). *Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning*.

Mokhtari, S., Yen, K. K., & Liu, J. (2021b). *Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning*.

Mukherjee, S., Sadhukhan, B., Sarkar, N., Roy, D., & De, S. (2023). Stock market prediction using deep learning algorithms. *CAAI Transactions on Intelligence Technology*, *8*(1), 82–94.

Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E., & Shahab, S. (2020). Deep learning for stock market prediction. *Entropy*, *22*(8).

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; A Comparative Analysis. *IEEE Access, 8*, 150199–150212.

Naik, N., & Mohan, B. R. (2019). Stock price movements classification using machine and deep learning techniques-the case study of indian stock market. *Communications in Computer and Information Science, 1000*, 445–452.

Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, *26*(4), 164–174.

Ravikumar, S., & Saraf, P. (2020). 2020 International Conference for Emerging Technology (INCET) : Belgaum, India. Jun 5-7, 2020. *2020 International Conference for Emerging Technology (INCET) : Belgaum, India. Jun 5-7, 2020.*, 1–5.

Sarode, S., Tolani, G. H., Kak, P., & C.S., L. (2019). Stock Price Prediction Using Machine Learning Technique. *ICISS-2019 : Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2019) : 21-22, February 2019*, 177–181.

Sen, J., Mehtab, S., & Nath, G. (2023). *Stock Price Prediction Using Deep Learning Models*.

Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of Big Data*, *7*(1).

Shilpa, B., & Shambhavi, B. (2023). Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments. *Kybernetes*, *52*(3), 748–773.

Singh, S., Kumar, J., Madan Kumar, T., & Singh Kumar, A. (2019). Stock Market Forecasting using Machine Learning: Today and Tomorrow. *2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (*, 738–745.

Sirimevan, N., Mamalgaha, I. G. U. H., Jayasekara, C., Mayuran, Y. S., & Jayawardena, C. (2019). Stock Market Prediction Using Machine Learning Techniques. *2019 International Conference on Advancements in Computing : ICAC*, 192–197.

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. In *International Journal of Financial Studies* (Vol. 11, Issue 3). Multidisciplinary Digital Publishing Institute (MDPI).

Yadav, A., Jha, C. K., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science*, *167*, 2091–2100.

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. *Neural Computing and Applications*, *32*(6), 1609–1628.

Zhao, Y., & Yang, G. (2023). Deep Learning-based Integrated Framework for stock price movement prediction. *Applied Soft Computing*, *133*.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, *5*(1).