Stock Market Prediction of Bangladesh Using Multivariate LSTM with Sentiment Identification

BY

Md. Ashraful Islam
ID: 191-15-12271

Md. Rana Sikder
ID: 191-15-12511

Sayed Mohammed Ishtiaq
ID: 191-15-12530

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Abdus Sattar

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By
Md. Sadekur Rahman

Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled "**Stock Market Prediction of Bangladesh Using Multivariate LSTM with Sentiment Identification**", submitted by **Md. Ashraful Islam**, **Md. Rana Sikder** and **Sayed Mohammed Ishtiaq** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on January 2022.

BOARD OF EXAMINERS

____
Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

24.01.23
____
Dr. Mohammad Shamsul Arefin — **Internal Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

24.01.23
____
Md. Sabab Zulfiker — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

24.1.2023
____
Dr. Ahmed Wasif Reza — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Abdus Sattar, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Md. Ashraful Islam
ID: 191-15-12271
Department of CSE
Daffodil International University

Md. Rana Sikder
ID: 191-15-12511
Department of CSE
Daffodil International University

Sayed Mohammed Ishtiaq
ID: 191-15-12530
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Abdus Sattar**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning & Data Science" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan,** Professor and Head, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

Stock market prediction is always challenging due to its volatile and dynamic movement. Apart from the technical factors, many external factors make it more difficult to predict the stock market of a developing country like Bangladesh. Therefore, it is not possible to accurately predict the stock market of Bangladesh by taking only the technical factors into consideration. Various studies have shown that some external factors like news sentiment, inflation, Gross Domestic Product (GDP), exchange rate, interest rate, and current balance of the country can affect the stock market trend, which is also applicable to Bangladesh. The main objective of this paper is to predict the trend of Dhaka Stock Exchange (DSEX), the largest stock market in Bangladesh by taking into account the technical stock market data as well as those appropriate external factors. This paper also compared the difference between the trend prediction with and without using news sentiment. All the technical and external stock market data from 2014 to 2021 is collected from verified sources. A multivariate Long Short-Term Memory (LSTM) neural network is used to predict the stock market trend. The experimental results indicate that news sentiment provides better performance in LSTM stock market trend prediction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The stock market plays an important role in keeping the wheels of a country's economy moving. Stock market prediction can therefore be considered an important study for the economy of a country. Although stock market prediction is a very difficult and complex process and there are always attempts to predict the stock market. But it becomes more difficult due to its dependence on various external factors [1]. The Bangladesh stock market is no exception. With digital advancements, stock market information has become more accessible and useful to investors. Now researchers can use more advanced tools and computers to make stock market predictions [2]. Every year researchers are releasing new methods and algorithms for stock market prediction. Machine learning and data analysis are widely used in this field as a new resource [3].

There are various technical and external factors on which stock market trends depend. While technical factors are easy to come by, external factors are very difficult to consider. Also, the stock market of developing countries like Bangladesh may depend on tougher factors for example news sentiment. Therefore, it is not possible to get accurate results in stock market prediction only by calculating the technical factors. In this paper, we have considered technical factors like - open, high, low, close, volume, and external factors like - news sentiment score, Inflation, GDP, Exchange rate, Interest rate, and Balance. The main goals of this work are as follows: i) To identify the correlation between Bangla news sentiment and stock market price trend, ii) To predict the stock market price trend using technical and external factors. The main objective of this paper is to predict the stock market of Bangladesh more accurately by considering external factors like Bangla news sentiment along with the stock market technical indicators.

1.2 Motivation

Stock market is an important part of our economy. Predicting the stock market trend is quite difficult in a country like Bangladesh where various external factors are making the stock market unpredictable. The main motivation came from the external factors of our country like – political and news sentiment, Inflation rate, Exchange rate, GDP are very much considered a root reason of sudden stock market ups and down. Various well-known algorithms are used in stock market prediction around the world. But they only consider the internal factors of the stock market. So, it is tough to predict the stock market of a country like Bangladesh.

But the main challenge of considering the external factors like – public market sentiment is tough because we cannot measure the sentimental position of a country with numbers. So, it would be great to find out a way to measure the market sentiment. We found national newspapers can be a great source for measuring the business sentiment for a stock market. If we get all the business news from any national news paper from predicted timeline than we can make a sentiment analysis from them and can use those data in our time series analysis phase. In Bangla language this news sentiment-based stock market prediction is completely a new approach. For that reason, we started to make a novel dataset for Dhaka stock exchange with sentiment column that will prove the necessity of considering business sentiment on stock market prediction.

1.3 Problem Definition

Predicting Dhaka stock exchange needs the external sentiment data because a country like Bangladesh stock market depends on external factors like – political situations, development, public sentiment, news sentiment, inflation rate, GDP, exchange rate, interest rate, current balance etc. For that reason, only internal factors are not enough for s tock market prediction in Bangladesh. But there is no such way to measure external public sentiment properly. In this thesis we proposed a way to measure the external public sentiment for our stock market. Daily business news is a great source of measuring the sentiment of the economic sentiment. So, we collected the daily business news from national newspaper by web scrapping. Then we did sentiment identification and generated a novel dataset with sentiment column for time series analysis.

1.4 Research Questions

Here are the main questions those are focuses in this thesis are given below:

- Is it possible to get better stock market trend prediction using external factor like News sentiment?

- How can we find a way to measure the public sentiment for stock market prediction in Bangladesh?
- Can we propose a better dataset for Dhaka stock market prediction with public news sentiment?

1.5 Research Methodology

In this section we described the Data collection process, Data Pre-processing, Sentiment Identification process for Bangla news sentiment analysis, LSTM model creation and Training the model. At the end the performance of the model and the outcomes will be described.

1.6 Research Objectives

In this paper, we have considered technical factors like - open, high, low, close, volume, and external factors like - news sentiment score, Inflation, GDP, Exchange rate, Interest rate, and Balance. The main objectives of this work are as follows:

- To identify the correlation between Bangla news sentiment and stock market price trend.
- To predict the stock market price trend using technical and external factors.
- To predict the stock market of Bangladesh more accurately by considering external factors like Bangla news sentiment along with the stock market technical indicators.

1.7 Report Layout

Chapter 1 will discuss about introduction, motivation, Problem Definition, Research Question, Research Methodology and the expected outcome of our project.

Chapter 2 will discuss about background of this research and the related work and works related to Bangladesh perspective.

Chapter 3 will describe the full research methodology.

Chapter 4 will discuss about our model performance data.

Chapter 5 it is focus to the result and correlation analysis of the outcomes.

Chapter 6 It describes the conclusion and future work of this research.

Chapter 7 here all the references we used for this research.

CHAPTER 2

BACKGROUND

2.1 Introduction

A lot of work has already been done on stock market prediction by considering technical indicators. We have found many works of stock market prediction taking technical factors into consideration in the context of Bangladesh as well.

2.2 Related Works

Technical indicators are not enough for predicting Bangladesh stock market. Reference [4] only used technical indicators as input and proposed a machine learning model which achieved almost 20% accuracy in stock market prediction. Reference [5] predicted the stock price index trend using an artificial neural network and support vector machine. They predicted the Istanbul Stock exchange using technical indicators. The linear classification model is also used to predict the stock market of Bangladesh [6]. This paper advised to consider Logistic regression to predict Bangladesh stock market.

The stock market can also be dependent on a country's GDP or economic growth [7]–[11]. Various papers described the dependence between a country's economic growth and the stock market. Economic growth or GDP can be an important external factor for predicting stock market trends [12]. Stock market return can also be affected by the inflation rate [13]–[16]. Stock market return price is positively correlated with the inflation rate.

News sentiment has a bigger aspect on stock market trends. Positive and negative news sentiment can play a vital role in building the stock market trend [17]–[20]. Reference [21] used the classifier Naive Bayes to detect positive and negative news sentiment and achieved up to 91.2% accuracy. In Bangladesh stock market, it depends on the Bangla financial news. So, we should consider Bangla financial news articles from some daily newspapers of Bangladesh. Social media sentiment is also used to predict the stock market trend. But it is always hard to get social media data for sentiment analysis. Also, the informal writing process in any social media platform makes it harder to achieve accuracy in sentiment analysis. Reference [22] used improved neural network algorithm to predict the Bangladesh stock market. They used external features for prediction.

The LSTM-based stock prediction method is more effective than other time series-based stock predictions. It can drastically increase the accuracy of a time series model. Supervised machine learning techniques with domain-specific Lexicon data dictionaries (LDD) give a better result on Bangla language text sentiment analysis. Reference [23] a rule-based BTSC algorithm that can classify the Bangla text sentiment.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Process Flow Diagram for the prediction and finding out the correlation.

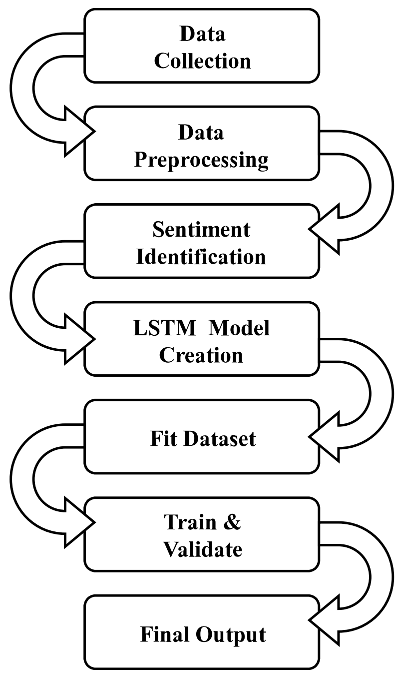

Figure 3.1 Process flow diagram.

3.2 Data Collection

For our multivariate LSTM neural network model, we used 11 variables as input. Technical variables: open, high, low, close, volume. News sentiment variable: news sentiment score. Economic variables: inflation, GDP, exchange rate, interest rate, and balance. So, our dataset CSV file has 11 variable columns. We can categorize these 11 variables in 3 parts for data collection:

1. Technical Dhaka stock exchange data from (3/3/2014 to 29/12/2021).
2. Financial news articles from The Daily Ittefaq (3/3/2014 to 29/12/2021) [24].
3. Yearly economic data from International Monetary Fund (IMF) database (2014 to 2021).

3.3 Technical stock data collection

We collected all the historical DSEX data from the Dhaka stock exchange's official website. The dataset has daily open, high, low, close and volume numbers of DSEX from 2014 to 2021.

3.4 Financial news articles for sentiment analysis

We collected 7695 financial Bangla news articles from The Daily Ittefaq using "Beautiful Soup 4.4.0" in python. It is a powerful web scraping tool for collecting public data from any website. We collected all the financial Bangla news articles from 3/3/2014 to 29/12/2021 [24]. There were multiple news articles on a single date. We collected and saved the news articles in CSV format. After collecting the news articles, we run our sentiment identification algorithm model and produced a sentiment score for every date using these news articles [23]. We created our news sentiment column in our dataset from these sentiment scores.

3.5 Yearly economic data

We collected the yearly economic data - inflation, GDP, exchange rate, interest rate, and current balance of Bangladesh from 2014 to 2021. We used IMF world economic database for this collection process. International Monetary Fund (IMF) is a verified source of economic data.

3.6 News sentiment identification

Our methodology for news sentiment identification came from this paper [23]. This paper proposed a rule-based algorithm for Bangla sentence-level news sentiment analysis called the BTSC algorithm. In Bangla language sentiment analysis, some words can behave differently according to the specific domain.

For example, in the phrase "শেয়ার ইস্যু করা", the word "ইস্যু" has a different meaning in business articles domain than our everyday life. So, to get an accurate result in sentiment analysis on financial Bangla news articles we have made our own business domain-specific extended sentimental dictionary dataset manually.

Our working steps are described below for Bangla news sentiment identification:

a) We constructed a finance domain-specific weighted sentiment lexicon dictionary in the Bangla language.

b) Used modified BTSC algorithm to analyze sentimental scores from Bangla financial news articles [23].

c) Then we collected 7695 financial Bangla news articles from the year 2014 to 2021.

d) After collecting the data, we run the sentiment analysis process and gathered sentiment scores for each day from 2014 to 2021 [24].

e) We merged these date-wise news sentiment data with our existing technical indicator dataset.

3.7 Creation of financial lexicon data dictionary (LDD)

The financial lexicon data dictionary is the list of words that are used to calculate the sentiment of financial news articles. We collected all the Bangla words from an online Bangla dictionary API and then classified them into 6 different weighted categories manually. Lexicon data dictionary is very important to find the exact sentiment report from the sentences.

All the words in our lexicon data dictionary are Bangla words. There are words considered as positive sentiment words and words considered as negative sentiment words. Some words have effects to increase the sentimental value of a sentence. So, we included those words to get proper sentiment visualization in Bangla. Some examples of our financial lexicon data dictionary (LDD) are shown in Table 3.1.

TABLE 3.1. FINANCIAL LEXICON DATA DICTIONARY

Types	**Examples**	**Weight**	**Word count**
Bull words	সচ্ছল, উন্নত, পরিকল্পনা, শীর্ষ, লাভবান, সমাধান, কার্যকর	+1	3385
Bear words	অনুপস্থিত, অকার্যকর, বাতিল, দুর্ঘটনা, দেউলিয়া, লঙ্ঘন	-1	3966
Negative words	না, নেই, নাই, নয়, নিষেধ, নও, ন	-1	35
Coordinating conjunction	বরং, কিন্তু, এবং, অথবা	+2	21
Subordinating conjunctions	বিশেষত, অধিকন্তু, এমনকি, এসত্তেও	+1.5	14
Adjectives & Adverbs	সবচাইতে, সর্বাধিক, যথেষ্ট	+3	23
	অধিক, বেশি, অনেক	+2	13
	সামান্য, প্রায়	+0.5	25

Bull words:

We named this wordlist bull words because they are categorized as positive sentiment words from a financial perspective.

Bear words:

Bear word list is the opposite of positive sentimental words in financial sentiment analysis. All the words in this list are considered as contradict sentiment words for analyzing the business news sentiment.

Negative words:

Negative word list has words like “না”, “নয়”, and “নেই” which can make a full sentence negative in the Bangla language.

Coordinating conjunction words (Co con.):

In the Bangla language conjunctions like “কিন্তু”, “আদপে”, “এবং”, “অথবা” plays an important role in sentence making. They should have their own weighted effect value in sentiment analysis [23].

Subordinating conjunctions (Sub con.):
Another kind of conjunctions list with words like “অধিকন্তু”, “এমনকি”, “বিশেষত”.

Adjectives & Adverbs (Adj.):
We listed some adjectives and adverbs like “সবচাইতে”, “অধিক”, “সর্বাধিক” as they are used to glorify the sentence sentiment more than other simple words. We categorized them into 3 weighted categories - high, medium, and low.

3.8 Data Tokenization

In this part, we preprocessed our data to make our news sentiment analysis. We got our data from our data collection step. We used BLTK [25] a python PyPI package that is OSI approved, License MIT to preprocess our dataset. Data preprocessing will help us to make our data more appropriate for our news sentiment identification model.

The first process of data preprocessing is data tokenization. Data tokenization is a method to split the whole news into sentences and then split all the sentences into words. We will do sentence-level sentiment analysis on our financial news articles.

For example: Here is a full news from our dataset - “টানা দরপতন আর লেনদেন খরায় চলছে শেয়ারবাজার। বেশ কিছুদিন ধরে ধারাবাহিকভাবে কমতে কমতে গতকাল দেশের প্রধান শেয়ারবাজার ঢাকা স্টক এক্সচেঞ্জের (ডিএসই) লেনদেন ৪০০ কোটি টাকার নিচে নেমে গেছে। গত বছরের ৫ এপ্রিলের পর ডিএসইতে ৪০০ কোটি টাকার কম লেনদেন হলো। এমন লেনদেন খরার বাজারে প্রায় ১০০ প্রতিষ্ঠানের ক্রয়াদেশের ঘর শূন্য হয়ে পড়ে। ফলে বড় পতন হয়েছে মূল্যসূচকের।”

After the sentence tokenizing process, it will be converted into - [‘টানা দরপতন আর লেনদেন খরায় চলছে শেয়ারবাজার।', 'বেশ কিছুদিন ধরে ধারাবাহিকভাবে কমতে কমতে গতকাল দেশের প্রধান শেয়ারবাজার ঢাকা স্টক এক্সচেঞ্জের (ডিএসই) লেনদেন ৪০০ কোটি টাকার নিচে নেমে গেছে।', 'গত বছরের ৫ এপ্রিলের পর ডিএসইতে ৪০০ কোটি টাকার কম লেনদেন হলো।', 'এমন লেনদেন খরার বাজারে প্রায় ১০০ প্রতিষ্ঠানের ক্রয়াদেশের ঘর শূন্য হয়ে পড়ে।', 'ফলে বড় পতন হয়েছে মূল্যসূচকের।’]

Then we will run word level tokenizing in every sentence. The first sentence will be like this –

['টানা', 'দরপতন', আর', 'লেনদেন', 'খরায়', 'চলছে', 'শেয়ারবাজার', ' ।’]

3.9 Data Normalization

Data normalization is the process of removing the characters which are not necessary for sentiment analysis. Characters like “,”, “।”, “;”, “#”, “!”, “@”, “%”, “$” have no meaning in sentiment counting. We removed all these special characters from our tokenized datasets. We also removed question marks. We skipped the process of considering question marks in sentiment scores like the BTSC algorithm [23].

3.10 Stop words

We removed various stop words from our normalized data. Stop words are considered as zero impact words on the sentiment dataset. For example, words like “জন্য”, “পর্যন্ত”, “নাগাদ”, “নিতে”, “হত” have no impact on sentiment analysis. We used BLTK stop word list to remove stop words from our dataset [25]. We used the level hard of the BLTK package for stop word removing.

3.11 Data stemming

Data stemming means to remove any suffix/prefix of a word and convert the word to its root form. We will remove parts like - “টি”, “টার”, “গুলো” and “গুলি” these types of suffix and prefixes from the word. For example: “শেয়ারবাজারে” will be “শেয়ারবাজার”, “বছরের” will be “বছর” after stemming the word. Stemming the word will help to match our input data with our LDD dataset words to analyze the sentiment score.

3.12 Parts of speech tagger (POS tagger)

POS tagger means categorizing the data in their parts of speech form. It’s very important to detect the parts of speed because we have adjectives, adverbs, and conjunction lists in our LDD. We have to detect them to implement the algorithm of news sentiment analysis. In the Bangla language, there are 5 basic parts of speeches - বিশেষ্য, বিশেষণ, সর্বনাম, ক্রিয়াপদ, অব্যয় [23]. Among them সর্বনাম has no impact on sentiment score. We removed them in the stop word normalization stage.

3.13 Sentiment score counting

We have modified the BTSC algorithm in a certain way to use in our financial Bangla news sentiment identification domain [23].

In step 1 we run our first loop to read news CSV files from our dataset folder. In our dataset folder, 7095 news articles are in CSV format. Here we also declared and initialized a variable "News_score" equal to zero. In step 3 we run our second loop to tokenize each news at the sentence level. That means the whole news is tokenized in sentences.

After sentence level tokenizing we normalized the sentences and remove all the stop words from them. Here we also declared and initialized a variable "Sentence_score" equal to zero. Then we tokenized each sentence at the word level after running the third loop. In this loop, we basically split the sentence into words. Then we run the stemming function and POS tagger function in every word [25].

After that in step 10, we scanned our LDD word lists. In step 11 we checked if the word is a bull word. We added a score +1 with the sentence score if the word detected as a bull word. Same as the bull words we also checked bear words, conjunctions, adjectives, and adverbs from our LDD lists and gave the word its score if matched. The coordinating conjunction score is +2, subordination conjunction score is +1.5. Adjective and adverb - high, medium, low are +3, +2 & +0.5. In step 25 we checked if the word is in our negative words list. If matched we check the existing sentence score. The pseudocode is given below in Table 3.2.

TABLE 3.2. MODIFIED BTSC ALGORITHM

Steps:	Algorithm Pseudocode
1.	**for** each News_articles.csv in datafolder **do**
2.	News_score = 0
3.	**for** each Tokenize(Sentence) in News **do**
4.	Normalizing(Sentence)

```
5.          Stop word (Sentence)
6.          Sentence_score = 0
7.          for each Tokenize(Word) in Sentence do
8.             Stemming(Word)
9.             POS tagger(word)
10.            Scanning LDD
11.            if word is Bull word in LDD then
12.                   Sentence_score =+1
13.            else if word is Bear word in LDD then
14.                   Sentence_score =-1
15.            else if word is Co con. in LDD then
16.                   Sentence_score =+2
17.            else if word is Sub con. in LDD then
18.                   Sentence_score =+1.5
19.            else if word is Adj. (High) in LDD then
20.                   Sentence_score =+3
21.            else if word is Adj. (Med) in LDD then
22.                   Sentence_score =+2
23.            else if word is Adj. (Low) in LDD then
24.                   Sentence_score =+0.5
25.            else if word is Negative in LDD then
26.                   if Sentence_score >=0
27.                     Sentence_score *-1
28.         end for
29.     News_score =+ Sentence_score
30.     end for
31.  end for
```

If the score is positive, we multiply the score by -1. Because if the sentence score is already negative and we also multiply it with -1, the score will be a positive sentiment score which is wrong. After that the loop ends in step 28. We count each single full news score by adding all the sentences sentiment scores in step 29. In steps 30 and 31 others for loops end.

3.14 Simulation of sentiment score counting

Now we will demonstrate the simulation of our modified BTSC algorithm with an example: First let’s see the news - টানা দরপতন আর লেনদেন খরায় চলছে শেয়ারবাজার। ধারাবাহিকভাবে কমতে কমতে গতকাল শেয়ারবাজার ৪০০ কোটি টাকার নিচে নেমে গেছে। এমন লেনদেন খরার বাজারে প্রায় ১০০ প্রতিষ্ঠান ক্রয়াদেশ দিতে পারছে না । ফলে সর্বাধিক পতন হয়েছে মূল্যসূচকের।

After sentence level tokenization: ['টানা দরপতন আর লেনদেন খরায় চলছে শেয়ারবাজার।', 'ধারাবাহিকভাবে কমতে কমতে গতকাল শেয়ারবাজার ৪০০ কোটি টাকার নিচে নেমে গেছে।', 'এমন লেনদেন খরার বাজারে প্রায় ১০০ প্রতিষ্ঠান ক্রয়াদেশ দিতে পারছে না ।', 'ফলে সর্বাধিক পতন হয়েছে মূল্যসূচকের।']

After word tokenization, normalization, removing stop words & stemming:

1st sentence: ['টানা', 'দরপতন', 'লেনদেন', 'খরা', 'চলছে', 'শেয়ারবাজার']

2nd sentence: ['ধারাবাহিকভাবে', 'কমতে', 'কমতে', 'শেয়ারবাজার', 'কোটি', 'টাকা', 'নিচে', 'নেমে']

3rd sentence: ['লেনদেন', 'খরা', 'বাজার', 'প্রতিষ্ঠান', 'ক্রয়াদেশ', 'পারছে','না']

4th sentence: ['সর্বাধিক', 'পতন', 'হয়েছে', 'মূল্যসূচক']

Simulation of our modified BTSC algorithm is shown in Table 3.3. Here the news has a negative score that means the news has negative sentiment.

TABLE 3.3. SIMULATION OF MODIFIED BTSC ALGORITHM

Sentence	**Scoring**	**Sentence score**
1	টানা দরপতন লেনদেন খরা চলছে শেয়ারবাজার null bear bull bear null null 0 -1 +1 -1 0 0	-1
2	ধারাবাহিকভাবে কমতে কমতে শেয়ারবাজার কোটি টাকা নিচে নেমে bull bear bear null null null bear null +1 -1 -1 0 0 0 -1 0	-2
3	লেনদেন খরা বাজার প্রতিষ্ঠান ক্রয়াদেশ পারছে না bull bear null null bull null neg +1 -1 0 0 +1 0 *-1	-1
4	সর্বাধিক পতন হয়েছে মূল্যসূচক adj. (high) neg null null +3 *-1 0 0	-3
	News_score	-7

Here (*-1) means multiplying existing sentence_score with (-1).

3.15 Date wise news sentiment score counting

We counted all the news scores for each day and divided the total score by the news articles count on that day. Then we get the average news sentiment score for each date.

3.16 Multivariate LSTM model

Recurrent neural networks (RNN) models can learn the patterns in a time series dataset making them an extraordinary method for predicting time series datasets. LSTM neural networks can learn more accurate time series patterns for their long and short-term memory architecture. Stock market prediction is a challenging process because multiple variables or factors can affect its trend. For this reason, we considered multivariate LSTM instead of the univariate LSTM model. The stock market depends on factors like - Technical factors and External factors. Bangladesh stock market is also heavily dependent on some external factors like - news sentiment score, inflation, GDP, exchange rate, interest rate, and balance. The multivariate LSTM model considers the correlation of news sentiment to predict a more accurate result.

3.17 Time series analysis using LSTM in python

In our LSTM neural network time series stock market prediction, we considered the Dhaka stock exchange dataset from 2014 to 2021. The covid-19 situation is also included in this time zone which makes it more challenging to predict the original trend. We trained our model with these datasets and then predicted the trend. The last date of our stock market data is 29 December 2021. We also predicted the stock price of 30 December 2021 using our LSTM model. Our multivariate LSTM neural network model prediction steps are below:

1) Reading the dataset.
2) Feature selection and scaling.
3) Data Cleaning and transforming.
4) Training the LSTM neural network.
5) Predict Next Day’s Price.

3.18 Prerequisites

For the model building and prediction process, we used Python 3 environment and necessary python packages. We installed and used the following standard packages: pandas, NumPy, math, and matplotlib. We used pandas for data frame activities and matplotlib for plotting the graphs and results.

3.19 Dhaka stock exchange Dataset

Our dataset CSV file contains the stock market data of the Dhaka stock exchange from 2014 to 2021. We considered technical variables like open, high, low, close, volume, and some external variables like news sentiment score, inflation, GDP, exchange rate, interest rate, and balance. Our output variable is "Open". Because we can easily evaluate the stock market trend from the daily opening price of a stock market. All other variables are used as our input variables.
All the input variables in our dataset have a direct correlation with our output variable. All input variables can affect the trend of the stock market. We have a total of 11 input variable columns in our dataset. One of them "Opening price" is also used as an output prediction variable.
Preview of our dataset is shown in Table 3.4.

TABLE 3.4. PREVIEW OF DSEX STOCK MARKET DATASET

Date	Open	High	Low	Close	Volume	Sentiment	Inflation	GDP	Exchange	Interest	Balance
2014-03-03	4697.3	4732.25	4685.53	4687.19	50542994	0.5	7.34	172.88	77.64	9.08	1.406
2014-03-04	4687.19	4728.12	4687.19	4703.88	46282176	-1.5	7.34	172.88	77.64	9.08	1.406
2014-03-05	4703.88	4717.24	4684.67	4697.54	35469576	5.0	7.34	172.88	77.64	9.08	1.406
2014-03-06	4697.54	4724.82	4690.5	4699.63	41292218	-0.6	7.34	172.88	77.64	9.08	1.406
2014-03-09	4699.63	4715.65	4687.2	4687.2	51064991	4.6	7.34	172.88	77.64	9.08	1.406
2014-03-10	4687.2	4715.74	4663.99	4665.57	40176268	8.3	7.34	172.88	77.64	9.08	1.406

3.20 Feature selection and scaling Data Tokenization

We have to make our data clean and scaled to get an accurate prediction from our model. The feature selection part is a process where we select our input features from our dataset and the output feature for prediction. So, we selected 11 input features: open, high, low, close, volume, news sentiment score, inflation, GDP, exchange rate, interest rate, and balance.

One output prediction feature - opening price. After feature selection, we have to scale the dataset. It is very important to increase the accuracy of the training model. It also helps to better the model training time for the prediction. Scaling is the process of converting all the input data to a standard value range. We used the MinMaxScaler approach from the scikit-learn packages to scale all the input data to a range of 0 to 1. After finishing the prediction steps, we unscaled the data again to get the original values from our dataset.

3.21 Data Cleaning and transforming

Data cleaning is the process of finding the missing values in the dataset and filling them with numerical values. In our stock market dataset, we don't have any missing values. We applied the neural network LSTM sliding windows approach in our dataset to train and validate the dataset. We divided the data into 2 parts - training and testing dataset. 80% of our data used a training dataset for our model and the rest 20% we used for the testing process.

3.22 Training the LSTM neural network

Our multivariate LSTM neural network model architecture has 4 layers:

1) The first layer is an LSTM layer. It takes our mini-batches from the sliding window process as input and returns the whole sequence.
2) The second LSTM layer again takes the returned sequence from the first layer as input.
3) The third layer is a dense layer that consists of 5 layers.
4) The last dense layer returns the predicted value.

Our model sequence length is 50. So, each mini-batch from the sliding window process consists of a matrix of 50 steps and 11 features. The input neuron size of the first layer is equal to the size of our mini-batch input data. The total input layer of our LSTM neural model consists of (50*11) 550 neurons. Our epoch count is 50.

3.23 Predict Next Day’s Price

After training the multivariate LSTM neural network, we have to forecast our next day's opening price using this model also. As our sequence length has 50-time steps, we have to give a minimum of 50 steps of the dataset for the window sliding process. The model will return us the predicted opening price for 1 day which is the next day after our dataset ended.

CHAPTER 4

PERFORMANCE OF THE PROPOSED MODEL

4.1 Model performance

We fitted the train and validation dataset in our LSTM model. We run the model with the training and validation data for 50 epochs. The same process was done again without the external variables for comparison. Our training and validation loss curve without using news sentiment is shown in Figure 4.1:

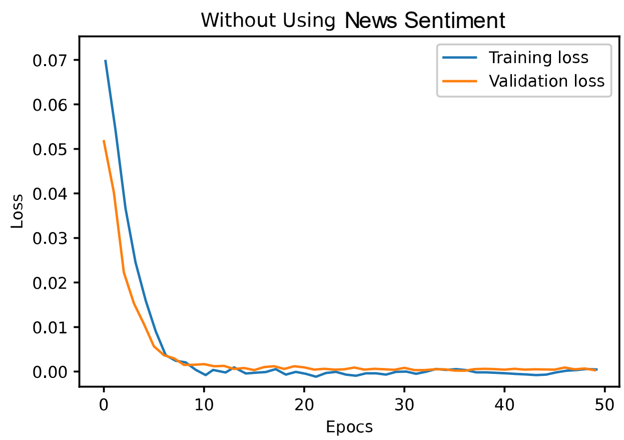

Figure 4.1 Loss curve without using news sentiment.

Our training and validation loss curve using all the 11 variables is shown in Figure 4.2:

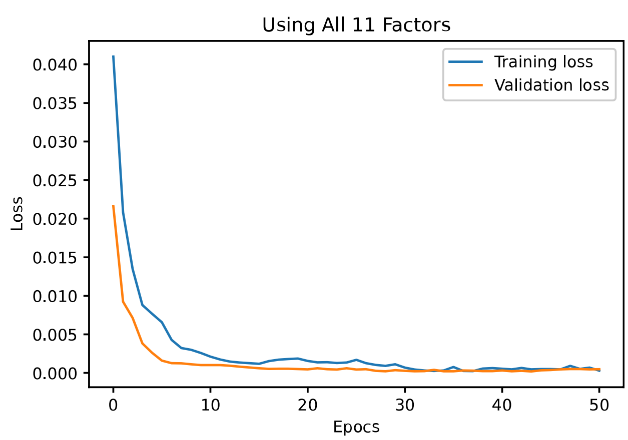

Figure 4.2. Loss curve using all the 11 factors

4.2 Comparison of the Accuracy

TABLE 4.1. COMPARISON OF ACCURACY (%)

Method	**Variables**	**Accuracy**
Multivariate LSTM without news sentiment	Date, Open, High, Low, Close, Volume, Inflation rate, GDP, Exchange rate, Interest rate, Current balance	63%
Multivariate LSTM using news sentiment	Date, Open, High, Low, Close, Volume, Inflation rate, GDP, Exchange rate, Interest rate, Current balance and News sentiment	72%

CHAPTER 5

RESULT COMPARISON AND ANALYSIS

First, we fitted and run our multivariate LSTM neural network model without news sentiment. We only used the variables for input data like - Date, Open, High, Low, Close, Volume, Inflation rate, GDP, Exchange rate, Interest rate, Current balance. Then we plotted the LSTM model predictions against the original opening price in Figure 5.1. The figure shows that the model without using news sentiment can only generate a long-term trend upward. But it fails to predict and generate a short time trend line which is not effective in daily trading. We have plotted only last few days for clear visualization.

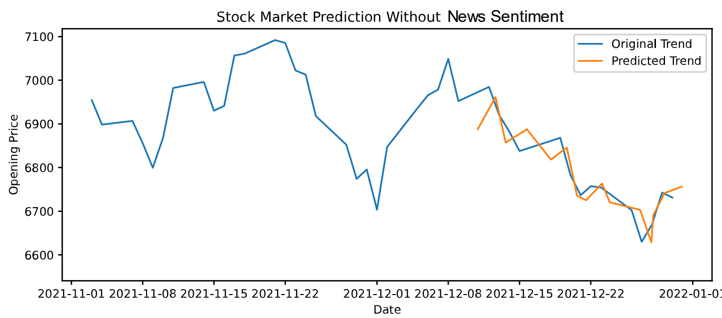

Figure 5.1. Opening price trend without news sentiment.

Then we fitted and run our LSTM mode with news sentiment including the technical and economic factors. A total of 11 variables were used as input. Then we again plotted the LSTM model predictions against the original opening price in Figure 5.2. We can see that the model predicts the stock market trend more accurately this time.

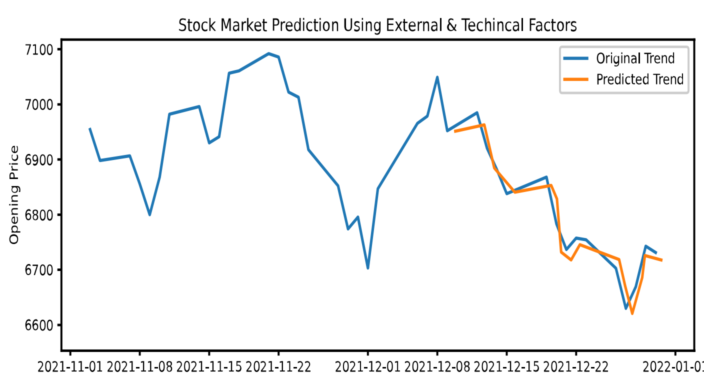

Figure 5.2. Opening price trend using news sentiment, technical & external factors.

We only measured the accuracy of this prediction graphically. Error calculation methods such as MAPE, MAE, RMSE will not be able to detect the accuracy of this result perfectly. So, we avoided the error calculation process intentionally. The model also predicted the opening price of the next day where our training dataset ended. It predicted the opening price of 30 December 2021 is 6721.46 taka. In real life the opening price was 6731.15 taka in Dhaka Stock Exchange.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The main objective of this paper was to prove the correlation between external factor like news sentiment for a better prediction of the stock market of Bangladesh. Only news sentiment alone has a huge capability of detecting many social affecting factors on the stock market trend. LSTM is a proven method of predicting the stock market but using multivariate LSTM makes it more reliable.

6.2 Future work

We can also use a Multivariate Multilayer LSTM or Bi-LSTM neural network for better prediction in the future. Multiple news sources can be used to collect the sentiment of a given time. Social media sentiment can also play a big role in this prediction. These external factors along with the financial news sentiment can produce better accuracy in multivariate LSTM stock market prediction.

REFERENCES

[1] M. Qi and G. S. Maddala, “Economic factors and the stock market: a new perspective,” *J Forecast*, vol. 18, no. 3, pp. 151–166, May 1999, doi: https://doi.org/10.1002/(SICI)1099-131X(199905)18:3<151::AID-FOR716>3.0.CO;2-V.

[2] I. Svalina, V. Galzina, R. Lujić, and G. Šimunović, “An adaptive network-based fuzzy inference system (ANFIS) for the forecasting: The case of close price indices,” *Expert Syst Appl*, vol. 40, no. 15, pp. 6055–6063, 2013, doi: https://doi.org/10.1016/j.eswa.2013.05.029.

[3] D. K. Sharma, H. S. Hota, K. Brown, and R. Handa, “Integration of genetic algorithm with artificial neural network for stock market forecasting,” *International Journal of System Assurance Engineering and Management*, 2021, doi: 10.1007/s13198-021-01209-5.

[4] G. S. Atsalakis and K. P. Valavanis, “Surveying stock market forecasting techniques – Part II: Soft computing methods,” *Expert Syst Appl*, vol. 36, no. 3, Part 2, pp. 5932–5941, 2009, doi: https://doi.org/10.1016/j.eswa.2008.07.006.

[5] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange,” *Expert Syst Appl*, vol. 38, no. 5, pp. 5311–5319, 2011, doi: https://doi.org/10.1016/j.eswa.2010.10.027.

[6] Md. Karimuzzaman, N. Islam, S. Afroz, and Md. M. Hossain, “Predicting Stock Market Price of Bangladesh: A Comparative Study of Linear Classification Models,” *Annals of Data Science*, vol. 8, no. 1, pp. 21–38, 2021, doi: 10.1007/s40745-020-00318-5.

[7] L. Carp, “Can Stock Market Development Boost Economic Growth? Empirical Evidence from Emerging Markets in Central and Eastern Europe,” *Procedia Economics and Finance*, vol. 3, pp. 438–444, 2012, doi: https://doi.org/10.1016/S2212-5671(12)00177-3.

[8] A. A. Enisan and A. O. Olufisayo, “Stock market development and economic growth: Evidence from seven sub-Sahara African countries,” *J Econ Bus*, vol. 61, no. 2, pp. 162–171, 2009, doi: https://doi.org/10.1016/j.jeconbus.2008.05.001.

[9] C. Ball and J. French, “Exploring what stock markets tell us about GDP in theory and practice,” *Research in Economics*, vol. 75, no. 4, pp. 330–344, 2021, doi: https://doi.org/10.1016/j.rie.2021.09.002.

[10] J. E. Stiglitz, “Capital Market Liberalization, Economic Growth, and Instability,” *World Dev*, vol. 28, no. 6, pp. 1075–1086, 2000, doi: https://doi.org/10.1016/S0305-750X(00)00006-1.

[11] C. Florackis, G. Giorgioni, A. Kostakis, and C. Milas, “On stock market illiquidity and real-time GDP growth,” *J Int Money Finance*, vol. 44, pp. 210–229, 2014, doi: https://doi.org/10.1016/j.jimonfin.2014.02.006.

[12] C.-K. Choong, A. Z. Baharumshah, Z. Yusop, and M. S. Habibullah, “Private capital flows, stock market and economic growth in developed and developing countries: A comparative analysis,” *Japan World Econ*, vol. 22, no. 2, pp. 107–117, 2010, doi: https://doi.org/10.1016/j.japwor.2009.07.001.

[13] M. Omran and J. Pointon, “Does the inflation rate affect the performance of the stock market? The case of Egypt,” *Emerging Markets Review*, vol. 2, no. 3, pp. 263–279, 2001, doi: https://doi.org/10.1016/S1566-0141(01)00020-6.

[14] A. Jain and P. C. Biswal, “Dynamic linkages among oil price, gold price, exchange rate, and stock market in India,” *Resources Policy*, vol. 49, pp. 179–185, 2016, doi: https://doi.org/10.1016/j.resourpol.2016.06.001.

[15] G. Bekaert and E. Engstrom, “Inflation and the stock market: Understanding the ‘Fed Model,’” *J Monet Econ*, vol. 57, no. 3, pp. 278–294, 2010, doi: https://doi.org/10.1016/j.jmoneco.2010.02.004.

[16] R. P. Pradhan, M. B. Arvin, and S. Bahmani, “Causal nexus between economic growth, inflation, and stock market development: The case of OECD countries,” *Global Finance Journal*, vol. 27, pp. 98–111, 2015, doi: https://doi.org/10.1016/j.gfj.2015.04.006.

[17] Y. Jiang, “Investor Sentiment and Stock Market Response to Earnings Announcement,” in *2011 International Conference on Management and Service Science*, 2011, pp. 1–4. doi: 10.1109/ICMSS.2011.5998465.

[18] X. Li, P. Wu, and W. Wang, “Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong,” *Inf Process Manag*, vol. 57, no. 5, p. 102212, 2020, doi: https://doi.org/10.1016/j.ipm.2020.102212.

[19] D. E. Allen, M. McAleer, and A. K. Singh, “Daily market news sentiment and stock prices,” *Appl Econ*, vol. 51, no. 30, pp. 3212–3235, Jun. 2019, doi: 10.1080/00036846.2018.1564115.

[20] M. W. Uhl, “Reuters Sentiment and Stock Returns,” *Journal of Behavioral Finance*, vol. 15, no. 4, pp. 287–298, Oct. 2014, doi: 10.1080/15427560.2014.967852.

[21] S. Kalra and J. S. Prasad, “Efficacy of News Sentiment for Stock Market Prediction,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 491–496. doi: 10.1109/COMITCon.2019.8862265.

[22] M. Billah, S. Waheed, and A. Hanifa, “Stock market prediction using an improved training algorithm of neural network,” in *2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, 2016, pp. 1–4. doi: 10.1109/ICECTE.2016.7879611.

[23] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. S. Islam, “Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon Dictionary,” *Natural Language Processing Research*, vol. 1, no. 3–4, pp. 34–45, 2021, doi: https://doi.org/10.2991/nlpr.d.210316.001.

[24] H. Tasmima, “The Daily Ittefaq.” https://www.ittefaq.com.bd/business (accessed Jun. 20, 2022).

[25] S. Hossain, “BLTK: The Bengali Natural Language Processing Toolkit.” https://pypi.org/project/bltk/ (accessed Jun. 15, 2022).

APPENDIX

The DSEX data collection phase proved to be a huge issue for us. All the data have been collected manually from DSE database server. For the news collection process, we used some cutting-edge new technology like web scrapping with beautifulsoup framework. Furthermore, creation of Bangla LDD was a difficult task. We used translated English LDD and refined it more manually.

Stock Market Prediction of Bangladesh

ORIGINALITY REPORT

8%	5%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.iki.bas.bg Internet Source	1%
4	Submitted to Liverpool John Moores University Student Paper	<1%
5	Submitted to Study Group Australia Student Paper	<1%
6	Razvan Zanc, Tudor Cioara, Ionut Anghel. "Forecasting Financial Markets using Deep Learning", 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), 2019 Publication	<1%
7	edissertations.nottingham.ac.uk Internet Source	<1%