UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF FINANCE – BANKING

————— ⌑ —————

GRADUATION THESIS

FORECASTING VN-INDEX TREND BASED ON ARTICLE HEADLINES AND TECHNICAL INDICATORS USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

Instructor: MBA. Phan Huy Tam

Student: Do Thi Lan Phuong

Student code: K194141741

Class: K19414C

Ho Chi Minh City, 04/2023

UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF FINANCE – BANKING

————— ⌑ —————

GRADUATION THESIS

FORECASTING VN-INDEX TREND BASED ON ARTICLE HEADLINES AND TECHNICAL INDICATORS USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

Instructor: MBA. Phan Huy Tam

Student: Do Thi Lan Phuong

Student code: K194141741

Class: K19414C

Ho Chi Minh City, 04/2023

ACKNOWLEDGEMENT

To complete this graduation thesis, the author has received the support and encouragement of many groups and individuals. First of all, the author would like to express the deepest gratitude to MBA. Phan Huy Tam (Instructor), who has always enthusiastically shared research experiences, professional suggestions, spiritual encouragement and answered questions throughout process of the study.

The author would like to thank the School Administrator of University of Economics and Law, the functional units of UEL for always creating favorable conditions for the author in the process of studying and researching. Especially, the Faculty of Finance and Banking, everyone's professional comments and thoughtful guidance are the motivation for the author to try in the future.

Finally, the author would like to express gratitude and appreciation for the love of family and friends who have always accompanied the author. Thanks to the spiritual encouragement, support and sharing of life, work and study experiences, the author has more motivation to be able to complete the graduation thesis.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATIONS

Bi-LSTM	Bidirectional Long Short-Term Memory
EMT	Efficient market theory
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
ReLU	Rectified Linear Unit
ROC	Rate of change
RSI	Relative Strength Index
SMA	Simple Moving Average
SVM	Support Vector Machine

ABSTRACT

The use of Machine Learning and Deep Learning techniques to forecast trend of stock market index is not too strange. However, to make forecasting models stronger in the short term, an approach studied in recent times is to combine sentiment analysis based on article headlines, historical time series and technical indicators. Most trading simulations performed in the financial sector are based on trend-following strategies that aim to identify and follow an ongoing price trend that has ability to persist in the following days. On the contrary, it is becoming quite common to apply Machine Learning and Deep Learning algorithms to predict trend reversals like changes in price direction. This study will apply two Machine Learning models including Logistic Regression, KNN and Deep Learning model - Bi-LSTM to forecast the trend of the VN-Index with the input factors which are sentiment analysis score of article headlines and technical indicators. Then, compare the performance of different classification models to choose the model that best fits the dataset and the best forecasting time period. The results of the study show that KNN gives the highest average accuracy in three models when combined with the analysis of the article headlines and technical indicators, while the Bi-LSTM achieves the highest accuracy when only using the input value is the score of article title analysis.

Keywords: Deep Learning, forecasting trend, Machine Learning, sentiment analysis, technical indicators

1. INTRODUCTION

1.1 Reasons for choosing the research topic

The financial market is considered the heart of the modern economy because it provides the means to sell and buy assets such as bonds, stocks, foreign exchange and derivatives. In particular, the stock market is one of the most important factors forming the economy of each country. Hafer and Hein (2007) asserted that without the stock market and the development of financial markets, there would be no significant growth in a country's economy. Although only formed and developed for less than 30 years, Vietnamese stock market has made progress when comparing to other countries in the region. Despite many ups and downs, Vietnamese stock market is increasingly proving its strength in terms of structure, scale and driving force to promote domestic economic development and international integration. Currently, on both HoSE (Ho Chi Minh City Stock Exchange) and HNX (Hanoi Stock Exchange), the number of stocks representing listed companies has surpassed 700. Of which, according to data obtained from the Vietnam Securities Depository (VSD), until 31/12/2022, the number of newly opened accounts reached 2.6 million - a record number in 22 operating years of the stock market. Therefore, Vietnamese stock market increasingly attracts the participation, attention of domestic and foreign investors. However, the stock market is a complex and unpredictable environment, investors need to have a firm grasp of information about businesses as well as carefully research historical stock prices to make effective and sustainable investment decisions.

Currently, in Vietnam, there have been many related studies on predicting stock price trends or market index including VN-Index/VN30-Index/HNX-Index such as the studies of Nguyen and Nguyen (2013), Truong (2014)... However, these studies mainly study and measure the influence of macro factors with basic criterias such as inflation, exchange rate, gold price, money supply and using traditional models such as ARIMA, VAR... Most of these reports are focused on mid-term and long-term goals. Previous researches on this topic had generally been divided into two main approaches: technical analysis and fundamental analysis. With technical analysis, mathematics has been widely used to analyze historical stock price patterns and predict stock prices in the near future. Meanwhile, fundamental analysis mainly looks at the intrinsic value of the stock price, financial indicators, business performance of the business and these metrics will usually be calculated for a specific time like a quarter or a year. Meanwhile, news related to the business can immediately affect the stock price, but in terms of financial data, it will be reflected in the next quarter's financial statements. This proves that news has a clear impact on the stock market and for a developing market like the Vietnamese stock market, this effect is even clearer.

Environment of the stock market is described as highly volatile with many external factors that directly affect stock prices (such as historical price, supply and demand, financial

articles...), which makes the process of predicting profits or short-term price trends very difficult for investors. Beginning from their own emotional predictions and mostly influenced by unclear short-term information, investors have ineffective stock transactions. Daily news events such as developing political situation, corporate activity, market and other unexpected socio-economic events affect the stock price immediately in a positive or negative direction. Therefore, it is not possible to accurately predict stock prices and their trends (up or down), instead investors can only forecast upcoming short-term trends. Investors often evaluate a company's business activities and related information before making a decision to buy shares. The evaluation includes analyzing the company's quarterly earnings reports and paying attention to important news to avoid buying undervalued or high-risk stocks. However, both the publishing speed and the number of daily news outlets have grown tremendously over the past few years, which overwhelms investors' ability to scrutinize huge volumes of data. Therefore, a coordinated automatic decision support system is essential because it automatically assesses and makes predictions about upcoming stock trends. For example, if the price of a potential stock is predicted to "rise" tomorrow, an investor can either sell the stock they hold for a higher price or wait for the price to fall and buy more. According to Fama (1970) stated on the Efficient Market Theory that: "A market is considered to be efficient about information if the price at that moment accurately reflects all available information." This was a widely accepted theory in the past. It is tied to Random walk theory, which argues that future stock prices are random and unpredictable. However, with the advent of technology, researchers have proven that stock prices can be predicted to a certain extent. One of the methods currently being researched quite a lot is sentiment analysis based on the title, content of articles and related news. Many scientists have used text mining tools and scored the article's sentiment according to the level of positive/negative/neutral to forecast stock prices and achieve high accuracy. Bollen et al. (2011) built the model with an accuracy of 87.6% when predicting the uptrend and downtrend of stock price trend from Twitter. Besides, according to Bujari, Armir, Marco Furini and Nicolas Laina (2017), historical market data combined with data extracted from social media platforms could be analyzed to predict changes in the economic and business sector. The performance of daily profit prediction systems depended heavily on the quality of the input values that the model was using.

With the advancement of technology, different models have been proposed both academically and practically for forecasting stock prices and market indexes. Previous studies on stock price prediction by Machine Learning models largely divided into two main approaches. The first approach aims to propose a predictive model using only historical stock data to represent for input values and the second approach applies related independent variables to create the model including external indicators (news sentiment and psychosocial) and technical indicators. Therefore, in this research paper, in order to continue to study more about Vietnamese stock market with short-term goal, the author uses two methods, namely

sentiment analysis and technical indicators with forecasting features to classify the VN-Index's up/down trend by daily return. In which, three models used as KNN, Logistic Regression and Bi-LSTM are applied to learn the relationship between independent and dependent variables.

1.2 Research objectives

General objective: The main objective of the report is to systematize theories and research papers related to market index trend prediction. Thereby, the author finds the gaps of the research articles after the review process. Research methods to exploit article headlines, calculate the sentiment scores and select appropriate technical indicators to help Machine Learning and Deep Learning models predict two trends of VN-Index with high accuracy to solve research gaps. From there, the author evaluates and compares the performance between the three models to discuss the effectiveness and suitability of each model in the trend classification problem and at different time periods in the short term. On that basis, personal investors are offered more reference sources to be able to make a decision whether to trade in the following days or not and make appropriate assessments for the stock market. Thereby, decide which industry to invest in, which stock code with an appropriate and accurate reference price.

Detailed objective: From the general goals in part 1.2, the author identifies the detailed goals that need to be implemented as follows:

1) Synthesize and apply theories related to VN-Index trend prediction. In addition, re-statistical gaps and limitations of the research articles reviewed.
2) Use available libraries to collect article headlines and techniques related to sentiment analysis to convert to sentiment score. And select appropriate technical indicators with research data.
3) Research and select two Machine Learning algorithms and one Deep Learning algorithm to build models. Then, compare and evaluate the exact effectiveness of the three models used. In addition, compare model performance when forecasting for different time periods in the short term.
4) Discuss the advantages as well as limitations of input values, Machine Learning and Deep Learning models. From there, recommend to investors more sources of reference on deciding whether to participate in investment or not. And if investing, how should it be invested in the short term?

Research question: After considering the aspects and issues that can be addressed from the topic of the research paper, the author poses the following questions:

Question 1: Which theories are used as a theoretical review for the prediction?

Question 2: Which libraries are used to analyze sentiment on the article headlines and which technical indicators are selected as input values?

Question 3: Which application model has the highest accuracy and high performance with what time period in the short term?
Question 4: Based on the results of the prediction models in the research, can investors refer to it to make decisions?

1.3 Research subject

The study focuses on the Uptrend and Downtrend trends of the VN-Index. In which, the prediction will be made by using Machine Learning and Deep Learning models based on article headlines and technical indicators. After building models and getting the results, compare the performance of the models to consider the effectiveness of trend prediction. The study evaluates the effectiveness of the model in predicting mainly through metrics such as Accuracy, Precision and Recall depending on each evaluation purpose. In which, the trend of VN-Index will be predicted according to the different time period focused timelines. Therefore, splitting the data set and predicting by time periods will make the prediction task of the model more objective and clearer.

1.4 Research scope

Regarding the spatial scope, the research is limited to the territory of Vietnam and uses financial data of Vietnam.

Regarding the time range, the study data is collected over a 5-year period from 01/01/2018 to 04/01/2023. This is a period consisting of 3 phases: before the Covid-19 epidemic, during the Covid-19 outbreak, and the recovery process after the Covid-19 epidemic when it has declined.

In particular, the dataset used for research, analysis, evaluation and building model will be collected from reliable and specialized information sites in the field of finance in general and securities in particular. At the same time, there are references from relevant documents and research articles with clear and scientific origin.

1.5 Research layout

Part 1 - Introduction: briefly state the reason of choosing the research topic to see the urgency to research, the research objectives including general and detailed objectives, research questions, research object, research scope and research layout.
Part 2 – Literature review: includes two parts: theoretical review and extant literature review. In which, the theory summarizes the concepts of efficient market theory, behavioral finance theory, information asymmetry, Machine Learning - Deep Learning algorithms including KNN, Logistic Regression, Bi-LSTM. Part 2 reviews previous studies to find research gaps.
Part 3 - The research methodology: clarify where the data comes from, summarize the independent and dependent variables. Meanwhile, the research method presents the implementation steps including data preparation, data processing, model building and model evaluation.

Part 4 - Research results: giving results of model evaluation based on metrics and specific cases according to each research purpose. In addition, compare models with each other to choose the most suitable model for each case.
Part 5 - Conclusion and Recommendations: systematizing all tasks done, answering research questions, presenting limitations of the study and proposing research directions in the future.

2. LITERATURE REVIEW

2.1 Theoretical review

In this section, the author will discuss the theories used in the research paper. Because this is a paper research on forecasting stock market indexes in the future, a number of related theories will be presented, including: Efficient market theory, Behavioral finance theory and Information asymmetry. Then there is the theory of predictive models built based on three algorithms: Logistic Regression, KNN and Bi-LSTM.

2.1.1 Efficient market theory

Efficient market theory (EMT) is one of the fundamental and important theories of the financial industry in general and the stock market in particular. This theory first appeared at the beginning of the 20th century, but it was not until 1970 that the researcher named Fama came up with the first concept. And after being agreed by many scientists, the concept of EMT is presented as follows: "A market will be considered to be effective in terms of information if the prices of commodities traded on the market are reflected timely and complete by the relevant information available". In other words, stock prices are reflected by investors' beliefs about future expectations. In particular, the relevant information available includes many different categories, such as information about the business performance and competitors of the business, information about the micro and macro economy. According to Robert (1967) and Fama (1970) have synthesized and classified the efficiency of the market into three forms as follows:

1) Weak form efficient market: the information set used at this level includes historical data of securities including prices, yields, ... in the past. If an investor can beat the market using historical stock price data or technical indicators, the stock market is not considered efficient.
2) Semi-strong form efficient market: the information set used at this level includes current and past published data, which includes weak form dataset. If an investor can beat the market using historical data along with public data such as news, earnings, and other stock fundamentals, the market is not considered efficient.
3) Strong form efficient market: the information set used at this level includes all known data, regardless of whether the data is published, this information set includes both weak and semi-strong form data sets. If investors have been beating the market using data from the first and second levels with personal information, then it can be concluded that the stock market is not considered highly efficient.

In addition, in order to apply EMT concepts to this study, it can be concluded that the first two levels of EMT are widely used by considering both historical data and public data for market index trend forecasting. However, the level with strong form efficient market, personal information is illegal in trading; therefore, it is not covered in this study. As stated above, the EMT states that the market fully reflects all available relevant information and

prices are fully and immediately adjusted as new information becomes available. If this is true then there will not be any benefit to the prediction, because the market will react and compensate for any action taken from this available information. In the real market, some people react to information as soon as they receive it while others wait for confirmation. Those who wait do not react until a trend is clearly established. Therefore, for short-term trend forecasting studies, typically technical analysis is seen as going against EMT. In fact, even price movements in the US and Japanese stock markets have been shown to follow only the weak form of EMT according to J. S. Ang & R. A. Pohlman (1978).

2.1.2 Behavioral finance theory

If Efficient market theory is stated that “The market is considered to be information efficient if the price at that time accurately reflects all available information” and the analysis of stocks is considered that prediction does not bring any benefit, behavioral finance theory with the fundamental core "the market is not always right" is considered as a great counterweight to EMT. EMT believes that once there is a market mispricing, there will exist an arbitrage opportunity and the investor thinks this is a reasonable opportunity to buy the undervalued asset and resell it at a higher price. This process will take place continuously, helping to adjust the market to equilibrium. Behavioral finance argues that this adjustment mechanism is not always possible. Unreasonable actions of investors occur continuously, becoming a trend in the market will lead to "price bubbles" and there will be a market crash. The terminology of behavioral finance theory first appeared in the 1930s and 1940s but is still limited. George Charles Selden published his work "Psychology of the Stock Market" in 1912, which was considered one of the world's first applications of psychology into the economy. Then, with the background research of Amos Tversky & Daniel Kahneman (1979), Richard H. Thaler (1985) and especially all the research of Robert Shiller (2000) through the book "Irrational Exuberance" laid a solid foundation for further developments in behavioral finance.

Behavioral finance is an important part of behavioral economics when it mentions to how psychological influences and biases affect each investor's decision. The theory has applied ideas that fall under the category of three main foundations, including: Psychology, Sociology and Finance. Psychology is the study of behaviors and thought processes to see how they are influenced by investor decisions by human thinking and the surrounding environment. Sociology will determine the social behavior of each individual rapidly and mainly focus on the influence of social relationships on human behavior and attitudes. Finance is the close concern between determining value and making investment decisions. Psychologists and other social scientists have studied human behavior for a long time and have accumulated considerable evidence on how people make decisions. With advanced technologies, neuroscientists might now show how brain activity can influence financial

decisions in the field of neuroeconomics. Instead of focusing on results, scientists focus on how human decision-makers achieve results.

The behavioral finance theory studied derives from the following three conditions on financial markets, including:

1) Exist unreasonable behaviors: some typical behaviors may follow such as irrational mental calculation, narrow definition, conservative, …
2) Systematic irrational behavior: the deviation in financial behavior is quite common for many individual investors, which has created a "herd effect" that makes stock prices not reflect true values.
3) Limiting the possibility of arbitrage in the financial market: as mentioned above, EMT believes that there will be an arbitrage opportunity if there is a mispricing. However, the cost of implementing arbitrage strategies and the trading existence of investors precluded this mechanism. The adjustments mentioned in EMT do not have an instantaneous adjustment in reality, but they often last for many years, so this is considered a sign of the limit of arbitrage possibilities.

Applying behavioral finance theory to explain individual investor behavior on the Vietnamese stock market is not still strange. This behavior can be explained by fulcrum-based investment psychology. In these cases, individual investors consider the world stock index as a fulcrum, a reference point to make predictions about the Vietnamese stock index. For example, when investors watch the increase of the Dow Jones index today, they will think that tomorrow, or the next few days, VN-Index will also increase, so they will make a decision to buy stocks, despite the fact that there is no scientific basis for the relationship between the variables in the same direction. Another type of behavior that is used quite a lot is investment decisions based on revealed information. This information can be from people around, from news sources,... Thus, the news can both serve as a reference source to decide and if the trend of the VN-Index is forecasted, it will also help investors find a fulcrum to decide their investment behavior.

2.1.3 Information asymmetry

In the market economy, information is understood as news that reflects new, current events and data that has significant value to the recipient. The value of information creates for people a disparity in benefits between not having or receiving information and the benefits of having received and used it. The quality and exploiting capabilities of the receiver will judge the value of the information. The information is more accurate, complete, and timely for decision-making, its value is said to be proportional to that decision higher. The value of information depends on many characteristics associated with the development and reality of the market, typically its hotness, confidentiality, difficulty to collect and its transmission. If the information is assessed to be complete, timely and accurate, it helps to make accurate forecasts, eliminate most of the risks in making decisions, exploiting investment opportunities

well is called perfect information. In practice, it is very rare that information is judged to be perfect because it is extremely difficult to collect, requires a lot of time and money and partly comes from the limited cognitive level of the information receiver.

In financial markets, one party often does not know enough information about the other party to be able to make correct decisions, this inequality is called information asymmetry. Information asymmetry occurs when buyers and sellers do not have access to the same information and sellers often have more information than buyers. For more than two decades, the theory of market with information asymmetry has been an important and urgent area of economic research. Today, countless applications extend from traditional agricultural markets in developing countries to modern financial markets in developed economies. The foundation of this theory was established in the 1970s by three researchers: George Akerlof, Michael Spence and Joseph Stiglitz. Considering the scope in the stock market when making stock transactions, information asymmetry reflects one or a group of subjects who own important information about the company that has not been disclosed to the public, while inaccessible to other investors (Chae, 2005). Information asymmetry can occur due to different sources of information, receiving information at different times, different ability to receive, understand, react and process information differently. This theory is both objective and subjective. Objectivity is considered to be due to the efficient level of the market leading to limitations on the transmission and updating information on prices. Subjectivity is commented on the reason that the lack of efforts as well as the lack of investment interest to search and exploit information of the subjects including individuals, organizations or businesses participating in the market.

The impact of information asymmetry performs in two ways: adverse selection occurs before the contract and moral hazard occurs after the contract (Santos et al., 2007). When a transaction occurs, an informed investor will earn a share of the benefit corresponding to the loss incurred by an uninformed investor. This loss is known as the reverse selection cost (Copeland & Galai, 1983; Glosten & Milgrom, 1985). The second type of consequence caused by information asymmetry is moral hazard. Moral hazard occurs after a transaction occurs, when one party has an incentive to behave differently after an agreement is made between the parties involved. On the basis of the assumptions of Ackerlof (1970) and research of Klann (2009), we can say that the direct consequence of information asymmetry is that in a transaction, the party owning a lot of information will receive more benefits from the other party. Therefore, information asymmetry mostly occurs when there is a difference in the level of information among two or more actors of a contract (Cardoso et al., 2007; Pires, 2008). Information asymmetry can be studied and measured by various methods such as standard price comparison method (Venkatesh & Chiang, 1986; Lee, 1993; Huang & Stoll, 1996) and econometric methods (Glosten & Harris, 1988; George et al., 1991; Madhavan et al., 1997). Each method has its own advantages and disadvantages depending on the different usage

conditions of the researchers. With countries having developing stock market, the econometric model of George et al. (1991) based on the market indicator variable is often applied to measure information asymmetry.

In recent years, with the development and advancement of textual information research, the revealing of non-financial and financial information plays a very important role in forecasting trends and making investment decisions. Faced with a large amount of information from financial reports and news, recipients often do not know where to start to dig into the text and tend to have information asymmetry. Therefore, the research paper wants to reduce the information asymmetry among investors, so it used the method of sentiment analysis from the article titles - the simple method of receiving the majority.

2.1.4 Logistic Regression, KNN and Bi-LSTM

2.1.4.1 Logistic Regression

In Machine Learning, Logistic Regression belongs to Supervised learning algorithms. Logistic Regression model does the job of predicting a dependent variable by analyzing the relationship between one or more independent variables. Mainly, this algorithm is used in classification problems. This algorithm has been used to build a model to predict default in the research of Ohlson (1980). Initial research focused on categorizing companies with label defaulting or non-defaulting firms. Logistic Regression is a method of statistical analysis to predict the probability of occurrence for an observation by adjusting the data using a Logistic curve. From there, the model gives a binary outcome based on the probability of each class. With a threshold of 0.5 (default), cases with a probability higher than this threshold will be classified as 1, while cases below this threshold will be classified as 0. For example, a Logistics Regression can be used to predict whether a high school student will be admitted to a particular university. These binary outcomes allow simple decisions to be made between two selections. In stock price prediction, the next month's price trend can be classified into two categories. In which, '1' indicates the next month's price trend is up or unchanged from the current month, and '0' indicates a downtrend. Logistic Regression models with two or more explanatory variables are widely used in practice (Haines et al., 2007). The calculation formula of Logistic Regression is presented in formula 1:

$$\ln\left(\frac{p}{p-1}\right) = b_0 + b_1 x \qquad (1)$$

From this equation, the probability of the output value is calculated according to formula 2:

$$p = \frac{1}{1+e^{-y}} \qquad (2)$$

Where:

p is the probability running from 0 to 1

b_0 is the intercept

b_1 is the slope coefficient of the input variable x

x is the independent variable

Typically, Logistic Regression forecasting model will be built through a series of processes including selecting the independent variables, determining the dependent variable, training the model, choosing the optimal regression coefficient, suitable threshold with g-means and testing with real dataset.

2.1.4.2 KNN (K-Nearest Neighbors)

Among the algorithms belonging to the group of Supervised learning algorithms, KNN is considered to be the most basic and simplest classification technique when there is little or no knowledge about the data distribution (L Devroye, 1981; L Devroye et al., 1977, 1982, 1994). Therefore, KNN is evaluated as a lazy learning algorithm. The KNN algorithm assumes that similar data will exist close to each other in a space. Therefore, its task is to find the output of a new data based on the output of the K nearest points around it. The class (label) of a new data object can be predicted from the classes (labels) of its K nearest neighbors. In this method, each sample will be classified similarly to the surrounding samples. Therefore, if the classification of a sample cannot be determined, it can be predicted by considering the classification of the nearest neighbor samples. KNN algorithm can be described as follows:

1) Identify the parameter K - the number of nearest neighbors
2) Calculate the distance of the object to need being classified to all the objects in the training set
3) Choose K which has the smallest distance.
4) Check the list of classes with the shortest distance, count the number of each class and choose the classifier for the class that appears the most times.

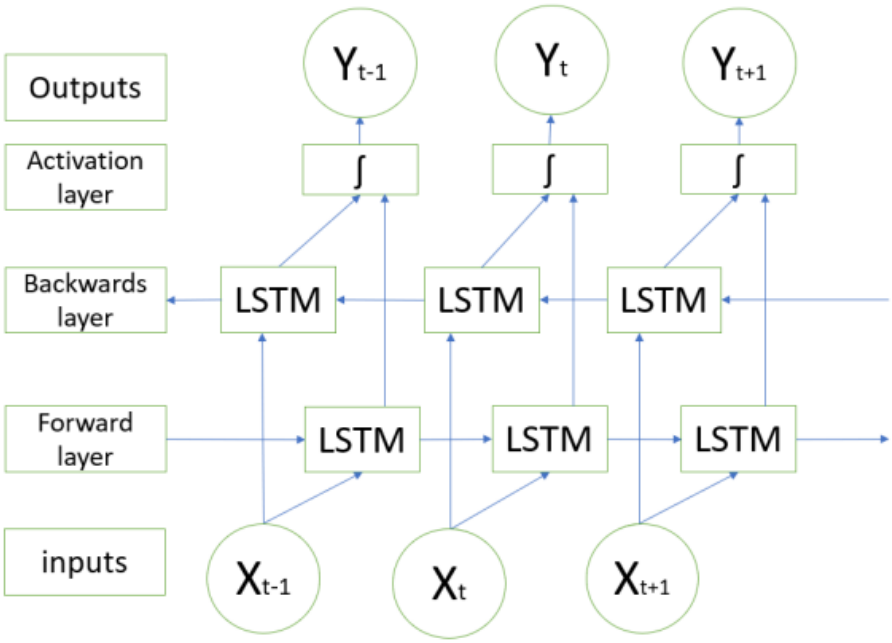
The two questions are mention to the model is really better if K is chosen as large as possible and how to choose the K that the model has the highest accuracy. The answer to question one is how it depends on the data. Larger K does not always give good results and vice versa. Answering the second question, the choice of parameter K of the model will be conducted through many experiments to choose the best result. Because the performance of the KNN classification model is mainly determined by the choice of the number K as well as the distance measure applied. In general, the affected classification results can refer to two aspects: the sparsity of the data and the noise points that will be mislabeled if K is too small and more outliers in the neighborhood than in other classes if K is too large. The biggest advantage of the KNN algorithm can be mentioned that it is computationally low in complexity, a simple algorithm for easy deployment and application as well as handles quite well with noisy data sets. The disadvantage of KNN is that it is easy to encounter noise factors and lead to inaccurate results with small K.

2.1.4.3 Bi-LSTM (Bidirectional Long Short-Term Memory)

With the rapid development of artificial intelligence, great advances have been made in neural network models over the past decades. Hochreiter and Schmidhuber (1997) proposed LSTM (Long Short-Term Memory) method to fix the problem that RNN (Recurrent

neural network) cannot process long sequence data due to vanishing or issues related gradient. The LSTM model has been widely used in many fields of science and engineering such as handwriting recognition (Graves et al., 2009) and stock analysis (Chen et al., 2015). The paper of Gers and Schmidhuber (2000) also found that stable sequences of peak values which are precisely timed and strongly non-linear can be generated using the LSTM technique. And it leads to a very promising method for real-world applications involving time and counts. Recently, tseveral variants of LSTM methods such as Stacked-LSTM (Du et al., 2017), CNN-LSTM (Huang and Kuo, 2018), Conv-LSTM (Liu et al., 2017) has been developed for different chronological learning. The prediction for a categorical variable in the future depends not only on the forward information under consideration but also on the back information as well. However, a layered structure of a traditional LSTM model with a single class can only predict the label of the current data based on the information obtained from the previous data. Bidirectional LSTM was created to fix the above weakness. A layered structure of Bi-LSTM usually contains two single LSTM networks and is used simultaneously and independently to model the input sequence in two directions: left-to-right (forward LSTM) and right-to-left (backward LSTM) as shown in Figure 1.

Figure 1. Structure of Bi-LSTM



Source: Reference from Sunny et al. (2020)

Bi-LSTM is a method of adding an backward LSTM layer to an current forward LSTM layer so that the hidden state will eventually create a hidden state connection vector between the two forward and backward LSTM layers. This inverted vector captures the hidden attributes and data samples that have been omitted in the LSTM layers. This allows the surrounding information for each hidden state to be equalized in both directions. Thereby, it

helps to solve the problem that the current LSTM encounters the problem of lost information. Bi-LSTM uses two layers of LSTM and is implemented by adjusting the data order of the layers. The first layer analyzes the data sample in the same forward direction as the LSTM, while the second layer analyzes the data sample in the opposite direction. The overall results are analyzed by linking, adding or averaging the analyzed results in two directions. The Bi-LSTM model is expressed in mathematical form throughout formula 3 and 4:

$$h_t^f = \tanh\left(W_{xh}^f x_t + W_{hh}^f h_{t-1}^f + b_h^f\right) \quad (3)$$

$$h_t^b = \tanh\left(W_{xh}^b x_t + W_{hh}^b h_{t-1}^b + b_h^b\right) \quad (4)$$

where: h_t^f and h_t^b refer to a front layer and a layer behind, respectively. W represents the weight connecting the input *x* to the hidden layer *h*. The hidden layer h_t is calculated by multiplying the input value x_t and the previous hidden layer h_{t-1} with each weight and replacing the value obtained by adding the vector bias *b* to the *tanh* function. The backup hidden class uses the value h_{t+1} instead of h_{t-1}.

The output value will be calculated according to the formula 5:

$$y_t = W_{hy}^f h_t^f + W_{hy}^b h_t^b + b_y \quad (5)$$

Clearly speaking, the output layer y_t is calculated by multiplying the weight W by the hidden layer in both directions and adding the bias *b*. By considering both hidden state values in both directions, the output value is output as a learned value not only for the forward propagation samples but also for the back propagation samples. The process of computing y_t by adding hidden states in both forward and backward directions through the above equation is shown as follows. The input values in each step are entered into the layers in both directions. The layers in both directions generate the h_t value through the LSTM information processing and the two h_t values are connected to the y_t output of each step. Bi-LSTM minimizes output loss and end-to-end learning is performed to learn all parameters simultaneously in both directions.

2.2 Extant literature review

Most of the previous studies have focused on textual or numeric input data and less looked at combining both types of data. However, with the development of advanced technology in the field of Machine Learning and Deep Learning, nowadays many researchers are interested in improving the performance of the model by using both types of data. Recently, RNN, LSTM or Bi-LSTM has gained great attention due to the rapid development of Deep Learning in the field of time series prediction thanks to the ability to resolve the long-term dependence of the data. Financial predictions based on Machine Learning often apply technical analysis to construct input variables. According to a survey by Atsalakis, G. S., & Valavanis, K. P. (2009), more than 20% of financial market prediction models use technical

indicators as input features. Therefore, many researchers have tried to demonstrate that media psychology affects stock prices and use it as an input value to create a predictive model.

With the huge amount of stock information, the investor's task becomes more difficult, because it has to collect, analyze, filter and make accurate decisions from this information. Information including historical financial information, real-time information and economic information suggest that in order to trade successfully in the financial market, it is important to develop models that one can identify different market states to modify their actions (Jan, Uzay & Willien, 2004). Technical analysts attempted to predict the stock market by learning charts that depict historical market prices and technical indicators (Suresh kumar and Elango 2011; Wei et al. 2011; Suthar et al. 2012; de Oliveira et al. 2013). In which, stock prices were preprocessed and appropriate indices were calculated to bring into the predictive model. Some technical indicators discussed in the study of Anbalagan and Maheswari (2014), Bisoi and Dash (2014) and Rajashree et al. (2014) were simple moving average (SMA), exponential moving average (EMA), moving average convergence divergence (MACD), relative strength index (RSI) and on-balance volume (OBV). Lamartine Almeida Teixeira et al. (2010) used four technical indicators as inputs for the prediction model including RSI, MA, Stochastics and Bollinger Bands. Meanwhile, Isaac Kof Nti et al (2020) combined both fundamental analysis and technical analysis as inputs. In which, SMA, EMA, MACD, OBV and RSI were used as technical indicators. A paper conducted by Taylan Kabbani et al. (2022) only selected three technical indicators to run the model including RSI, SMA and Stochastics. Thus, inheriting from previous studies, in this study, four technical indicators are used including SMA, RSI, ROC and Stochastics.

Tantisantiwong et al. (2020) proposed a system to forecast SET50 - a stock index in the Thailand market, using both data types including market data and textual information. For textual information, they collected social media data with advice from financial experts to identify sentiment analysis keywords (positive and negative). Then, observations would be labeled as positive or negative based on previously defined keywords. Finally, based on both numeric data (historical data of SET50) and text data (sentiment analysis), Tantisantiwong used multiple linear regression model to forecast the SET50 index. However, the limitation of this study is that these keywords are not widely published and they can be changed at different time periods so they must be updated periodically. Kingma et al (2014) developed a model to predict daily and monthly stock price movements using historical price data and sentiment analysis for the banking, mining and oil sectors. Historical prices were obtained from the website named yahoo.finance.com and textual dataset was generated using news and tweets over a year. Principal component analysis (PCA) with multiple factors was applied to the sparse dataset considered for evaluation sentiment analysis. In this study, three Machine Learning algorithms were used, Decision-Boosted Tree, SVM, and Logistic Regression to compare accuracy and use as performance measure. Decision-Boosted Tree algorithm went

beyond Logistic Regression and SVM when comparing with accuracy. Decision-Boosted Tree achieved 54.8%, 76% and 76.9% accuracy for the banking, mining and oil sectors, respectively. Logistic Regression achieved an accuracy of 65.4%, 61%, and 44.2% respectively and SVM achieved an accuracy of 51%, 59%, and 44.2% for the respective industries. This study recommended looking at the impact of intraday price movements on the next day's stock prices to improve accuracy. Through the comparison of accuracy, the author saw that the oil industry group was not predicted effectively in Logistic Regression and SVM models.

Vargas et al. (2017) presented the concept of collecting two different types of dataset including textual and numerical data. They collected news headlines and converted them into word vectors using the Word2Vec algorithm, then brought them to a Convolutional neural network (CNN) model to extract featured keywords. The obtained results with seven technical indicators calculated from the stock price history would be brought into the LSTM model as input features to predict the trend of the S&P 500 with the output variable like classified value. The results of testing showed that the highest accuracy is 64.21%. However, some limitations of this study are only forecasting for tomorrow, while that author's research objectives were suggested for weekly and monthly. The second is to use featured keyword extraction, but not analyze whether those keywords are negative or positive. And especially, this model is using too many independent variables, specially technical indicators that have no categorical significance, which made it difficult for the model to forecast. Li et al. (2020) studied the impact of different factors on the price volatility problem that investors could further apply in the stock market. The first is that the features were manually extracted from a certain oscillation period. The second is the usual technical indicators. And the third is used for denoising autoencoders (DAE) - a method capable of extracting Deep Learning features and removing noise from the time series input. This approach achieved an accuracy of 55.19% when it predicted bitcoin's price movements. For financial time series forecasting, Persio et al (2016) investigated the adequacy and proficiency in introducing the LSTM algorithm. Akita et al. (2016) compiled data using information from articles to show the influence of previous events on the opening prices of 50 stocks listed on the Tokyo Stock Exchange. They used the LSTM model to predict and concluded that there is efficiency when combining by industry. However, according to the research of Luca Di Persio (2017), the performance of Bi-LSTM model had better results than LSTM with MAE = 302.48 when the model was predicted short-term in the energy industry. Research of Md. Arif Istiake Sunny et al (2020) compared the performance of two algorithms LSTM and Bi-LSTM to predict stock prices. The results showed that Bi-LSTM gave higher efficiency when reaching RMSE = 0.0004127 with epoch = 30, while LSTM achieved RMSE = 0.0004928 with epoch = 100. However, this study still has one limitation, that is dividing the training set by random with the ratio of 88% and 12% respectively. Because the algorithm is applied for continous time series data with

the input values of indexes of historical price, the division of dataset is evaluated as inappropriate.

In the Vietnamese market, there have also been many studies on price prediction based on time series data and text mining from internet news, but there are still certain limitations. The article posted in Vietnam Trade and Industry Review (2021) by Truong T.T Duong used the LSTM model to predict the VN-Index and achieved quite good results with MSE = 13.8 and epoch = 10 in the training set, but did not mention the results in the testing set showing many inadequacies. Besides, the LSTM model is often used to predict for the short-term, but in this study, the input values are historical VN-Index values of the previous 60 sessions to forecast the next session. Meanwhile, the study of L. Minh Dang et al. (2018) deployed a good model when combining article titles and technical indicators to forecast the trend for the VN-Index with high accuracy. The highest accuracy reached 77.4% with the TGRU model. The small limitation in this study is that the article title is being restricted to a one-year period. Another study by Le Hong Hanh (2022) used a Machine Learning algorithm to analyze textual data from four famous news websites in Vietnam including Vietstock, Thanhnien, CafeF and Vnexpress to predict the VN-Index for the next day with types like increase, decrease and neutral. The four models used are Random Forest, Decision Tree, KNN and SVM. In which, the SVM model had the highest accuracy with 60.1% with the dataset belonging to Vietstock website. The final result is solid evidence that financial and securities news in the internet news and journals influences the price movements of VN-Index and the Vietnamese stock market.

After reviewing previous studies, the author finds out that the Vietnamese market still does not have too many models to forecast stock price trends or market indexes. There may be many research papers with forecasting purposes but still focusing on numeric data such as historical stock price or technical indicators or there is also research on sentiment analysis based on emotion throughout reviews, comments or article headlines, but there are still certain limitations. With the goal of building a model to forecast the stock market index trend in the short term, the author figures out that reading articles and considering technical indicators are also in the short term. Therefore, the study will proceed to build a predictive model that combines two types including textual data and numeric data. Besides, with the development and increasing innovation in technology, the author also wants to compare Machine Learning and Deep Learning models to see which model will be more effective in this prediction task.

3. RESEARCH METHODS

3.1 Data

3.1.1 Data source

The dataset in this study consists of two main sources. Firstly, for textual data, article collected from the website https://en.vietstock.vn include the title of the article and the publishing date in the 5-year period from 2018-2023 with 5267 observations. Secondly, for the historical data of VN-Index, this data is collected from TCBS and SSI by using the library "vnstock". Data were also obtained for a 5-year period from 2018-2023 with 1531 observations.

Table 1. Description of variables

Variable	Sign	Detail
Dependent variable	Trend	Uptrend (1)
		Downtrend (0)
Independent variable	Compound	Sentiment score
	Opening	Opening point of the day
	Closing	Closing point of the day
	Volume	Daily trading volume
	Lowest	Lowest point of the day
	Highest	Highest point of the day
	SMA	SMA Indicator
	RSI	RSI Indicator
	ROC	ROC Indicator
	%K	Stochastic Indicator

Source: Statistics of the author

In which, the input data is used to build the model including ten variables like Table 1: Sentiment Score (Compound), Daily Trading Volume (Volume), Highest point of the day (Highest), RSI Indicator (RSI), ROC indicator (ROC) and Stochastic Indicator (%K). With the dependent variable, the trend of VN-Index is a categorical variable with two classes: Uptrend and Downtrend encoded into two values of 0 and 1.

3.1.2 Technical Indicators

SMA (Simple Moving Average): This is a line calculated by averaging of range of the closing prices over a selected trading period. Calculated by formula 6 (with n = 14 trading days):

$$SMA = \frac{P_1 + P_2 + P_3 + \cdots + P_n}{n} \quad (6)$$

RSI (Relative Strength Index): A momentum indicator was developed by J. Welles Wilder, Jr. This indicator is used to identify when a stock is overbought or oversold within a specified period (in Here the author uses 14 trading days). The value of RSI ranges from 0 to100 and can be calculated from formula 7:

$$RSI = 100 - \frac{100}{1 + \frac{Average\ gain}{Average\ loss}} \quad (7)$$

%K: The Stochastic Oscillator is another momentum indicator that compares a stock's closing price to its price range over a period of time. It can be used to foretell a reversal when there is an increasing or decreasing divergence and constituted by two lines including %K and %D, where %K is the main line. The common timeframe used is 14 trading days and this indicator is calculated by formula 8:

$$\%K = 100 * \frac{P_C - LL_{14}}{HH_{14} - LL_{14}} \quad (8)$$

where: P_C is the closing price

LL_{14} is the lowest price in the past 14 days

HH_{14} is the highest price in the past 14 days

ROC (Rate of Change): This is an indicator used to measure the volatility of prices at two different times. The indicator will track changes in the market, helping investors to recognize the percentage change in the price, thereby helping to identify signs when falling into the overbought or oversold zone. ROC is calculated as formula 9:

$$ROC = \left(\frac{Current\ Price}{Precious\ Price} - 1\right) * 100 \quad (9)$$

3.1.3 Data labeling

The target of labeling in the study is to classify news that will reflect positively or negatively, leading to an uptrend or a downtrend of the VN-Index based on daily return. For this classification problem, there are generally three approaches to classifying trends by using closing and opening prices: close-to-close return (total daily return), open-to-close return (daytime return) and close-to-open (overnight return) according to Qingfu Liu & Yiuman Tse

(2017), Fengzhong Wang et al (2009). Inheriting from previous studies and practical viewpoint, with the reason that the records of the open-to-close are more similar to the total of the same stock and contributes to total daily return, the author will use the open-to- close approach to calculate daily return and label it as follows:

$$\text{Trend} = \begin{cases} Uptrend\ if\ R_i \geq 0 \\ Downtrend\ if\ R_i < 0 \end{cases}$$

Besides, to further assess the impact of different time period on VN-Index, the author considers different time periods (i.e., 1 day, 3 days and 9 days). The Open-to-Close return is calculated according to the formula10:

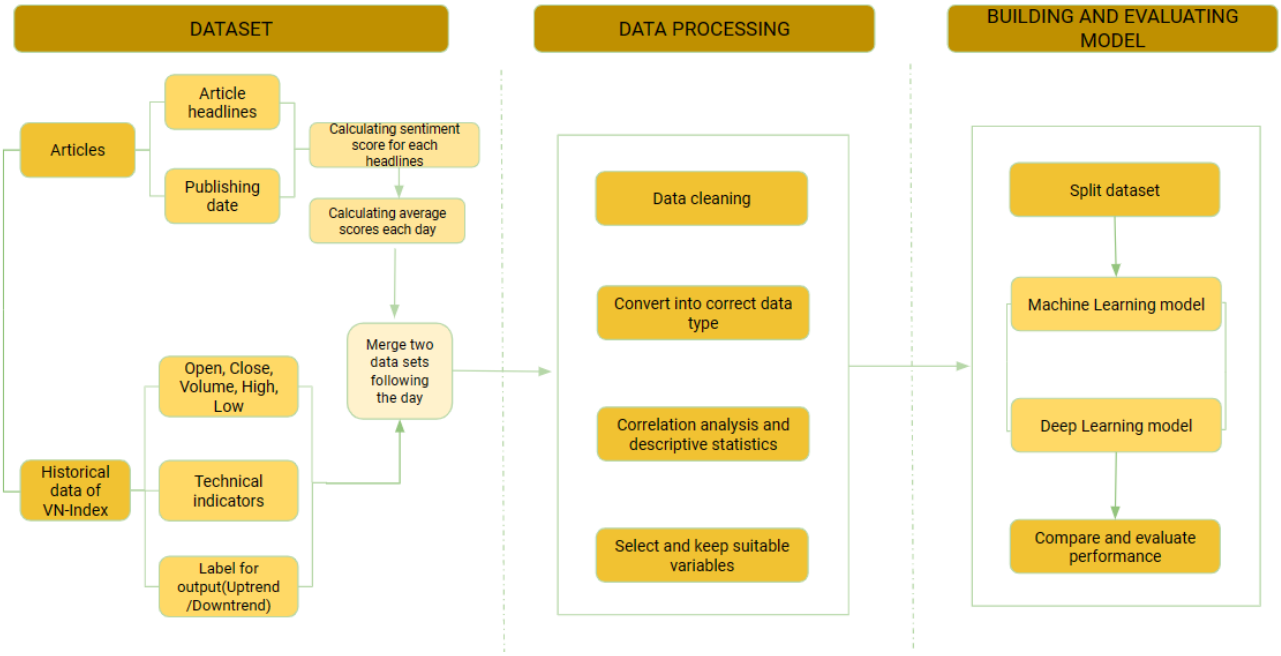
$$R_{dt} = O_{d+t} - C_d \qquad (10)$$

where: O_{d+t} is the opening price of the day after day d where t is the number of days ($t \in [1, 3, 9]$).

3.2 Research methods

The two main methods used in this research are sentiment analysis and technical analysis. Sentiment analysis, also known as opinion mining, is a field study about analyzing people's opinions, sentiments, evaluations, attitudes and feelings towards entities and properties shown in the text. This method is often used on textual data to test the attention, emotions, behaviors, decisions, and feelings of speakers or writers in relation to targeted topics. The term of sentiment analysis probably first appeared in the study of Nasukawa & Yi (2003) and the term of opinion mining first appeared in Dave et al (2003). The basic task in sentiment analysis is to group texts in a sentence or document to determine whether the text is positive, negative or neutral. Sentiment analysis techniques can be classified into three main approaches: lexicon-based approach, machine learning-based approach and hybrid approach. Meanwhile, technical analysis can be understood as a set of rules or charts that tend to predict future price changes based on the study of certain information such as closing prices, opening price, trading volume… (Gorgulho et al., 2011; Lin, Yang, & Song, 2011). Due to its sensitivity to historical data, technical analysis is often considered as an approach to invest in short-term and mid-term, it is also mainly used by short-term investors. This method was introduced in the 1800s by Charles Dow through Dow Theory. Technical analysis uses price patterns, such as candlestick/line charts and technical indicators to study and analyze stock price movements in the future. In which, technical indicators are indicators that are calculated based on stock price characteristics in the past such as closing price, opening price, highest price, lowest price or trading volume. From technical indicators, investors can identify price trends and reversal points.

Figure 2. An overview of market index trend forecasting process



Source: The author

As described in Figure 2, the proposed research process consists of three main parts including data preparation, data processing, building model and evaluating performance.

3.2.1 Data preparation

In the first step, there are two data sets that need to be prepared: the title of the article and the history value of the VN-Index.

Article headlines (Textual data): In order to ensure the effectiveness and popularity of the textual data along with the investor's demand for using online news sites, the author decide to use the data on the website https://en.vietstock.vn. The news on this website will mainly focus on the stock market, decisions of businesses, news of the domestic and international economy. The author uses the "Beautiful Soup" library and the "Request" library to crawl and collect data. From the news website of “vietstock”, the research paper only exploits to use the article headlines and publishing date because the author wants to hit the common sentiment of many investors when just reading through the title of article without carefully reading the full content of page. Then, the author convert the crawled data into a csv file, using the sentiment analysis tool named “Vader”, which is integrated into Natural Language Toolkit (nltk) library to calculate the sentiment score of each headline. This calculated score of each news ranges from -1 to 1. In which, closer to 1 indicates that the news is positive, closer to -1 means that the news is negative and if the score is close to 0, it expresses that the news is neutral. Finally, integrate and calculate the average score for each day. Before integration,

there are 5267 article titles collected and after integration, the data decreases into 1531 observations over 5 years.

Historical data of VN-Index (numeric data): To collect this data, the author uses the library named "vnstock". This library is developed by Thinh Vu (2022), which is a Python package to get Vietnamese stock market data from TCBS and SSI. “Vnstock” allows users to download historical stock data and market insights from TCBS. In which, the author uses features such as Closing, Opening, High, Low and Trading Volume to calculate technical indicators and as independent variables. Besides, Closing point, Opening point are also used to calculate and label for the output value. Extracting market data for a 5-year period from 2018-2023 has 1531 observations. To improve the accuracy of the classification model, the author calculates more technical indicators and implements labeling for the output values.

3.2.2 Data processing

In this step, there are four tasks to do:

1) Datatype conversion: Check and convert variables into the correct datatype. In which, with the two columns "Date" in the two data sets, both are converted to type of datetime before merging two data sets together. Check the datatype table and convert to the correct nature for each variable.
2) Data cleaning: Because when collecting historical price data and calculating technical indicators, null values and outliers will appear, so delete lines with values named “NaN”.
3) Descriptive statistics and correlation analysis: Check the correlation between variables, measure the dispersion of the variables and the statistics of the variables. This process will be presented in more detail in Part 4.
4) Select and keep the appropriate variables for the model: According to Pallant (2007): “Multicollinearity exists when the independent variables have a high correlation (r = 0.9 or more).” Therefore, the author will delete the pairs of variables if there is multicollinearity or the correlation is too high.

3.2.3 Building the models

Before building model, the author divides the dataset into two training/testing sets and normalizes the input data. After that, the author carry on training on the training dataset according to the algorithm including Logistic Regression, KNN and Bi-LSTM. The steps are as follows:

Step 1: Split the dataset

The dataset will be divided into two parts, one part for training and the other part for testing after training is complete. After merging and preprocessing the data, the study contained a total of 1246 observations. Because the data is a continuous time series, the author decided to use the first 900 observations as the training set and the rest of dataset with 346 observations as the testing set. Besides building a model to predict the trend of the VN-Index

with different time periods, the study will perform with three cases for each different input dataset.

Case 1: The independent variable is the sentiment score and the dependent variable is the trend of the VN-Index.

Case 2: The independent variables are technical indicators including RSI, ROC, Stochastic (%K), trading volume and highest value. The dependent variable is the trend of the VN-Index.

Case 3: The independent variable is a combination of the two input data sets of the two above cases and the dependent variable is the trend of the VN-Index.

Step 2: Standardizing the dataset

After data processing, dividing the dataset into training and testing sets, the author standardizes the data to bring the quantitative variables to the same scale to make data representation easier and the model Machine Learning/Deep Learning is more efficient. There are two main methods: Standardization (also known as z-score normalization) and Normalization (also known as min-max normalization). In this study, the author uses the Normalization method which is determined by the maximum and minimum values according to the formula 11:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (11)$$

where: x is the original value

x_{norm} is the normalized value

After being normalized, value of data will be in the range from 0 to 1. This might help balance the distribution of variables and avoid the phenomenon of features with large values.

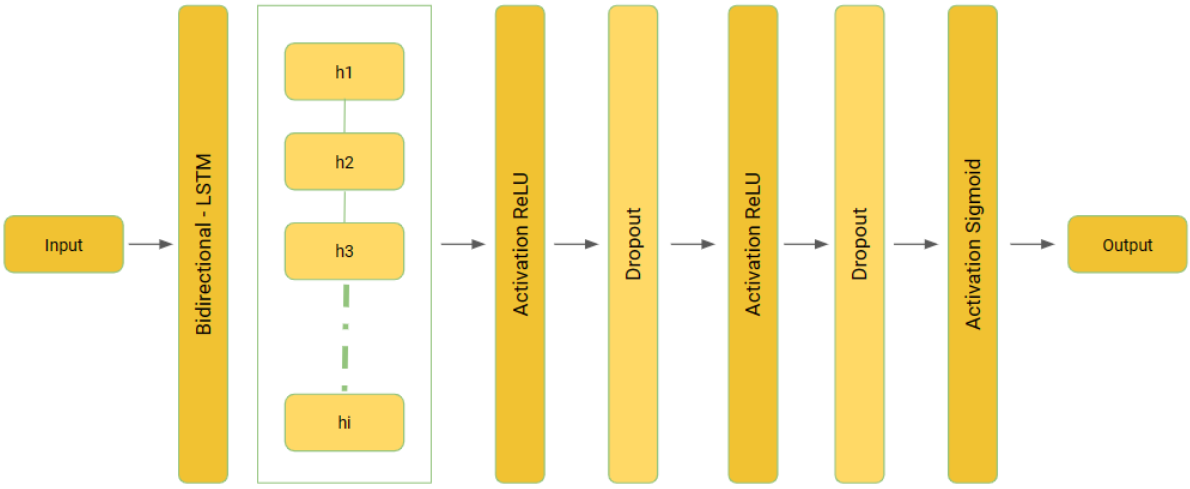
Step 3: Training the models

With the objective of comparing the efficient performance between models to choose the most suitable model, the author will build a training model with three algorithms Logistic Regression, KNN and Bi-LSTM.

1) Logistic Regression: Build a training model with a default threshold = 0.5. Then, test the model's accuracy to assess whether the performance is high or low. Next, calculate the g-mean score for each threshold to choose the threshold that best fits the model when the accuracy is high, the recall and precision indicators are not too different. After adjusting the model with the appropriate threshold, test it on the testing dataset to evaluate performance of the model.
2) KNN: Build a training model with default k = 5. Then, test the model's accuracy to evaluate whether the performance is high or low. Next, setup the graph showing between the two factors accuracy and “k”. From there, select the corresponding k with the highest accuracy when training with training dataset. After adjusting the model with the appropriate k, test on the testing dataset to evaluate performance of the model.

3) Bi-LSTM: Build a training model consisting of an input data layer with the number of data features of six, four hidden layers and an output layer with categorical variable, detailed as shown in Figure 3.

Figure 3. Structure of proposed Bi-LSTM model



Source: The author

The first hidden layer of the model contains 128 neurons. The two next hidden layers use activation function named Rectified Linear Unit (ReLU) containing 256 neurons ReLU function is said to be very effective in training Deep Learning models because it helps the model in training and optimizing the loss function faster (Krizhevsky, Sutskever, & Hinton, 2012). To reduce the phenomenon of overfitting, the author uses Dropout technique (dropout) with a rate of 0.2 after each hidden layer with the ReLU function. However, since the output data is a categorical variable, the final hidden layer will use the activation function named "Sigmoid". When combining layers together, the author uses a loss function with the attribute "binary_crossentropy" and the optimization function "adam". Besides, the author also uses the Early Stopping technique to automatically stop the training and save the weight value between layers that makes the loss function the smallest on the training set. In addition, in each iteration (epoch), because putting too much data in at the same time will lead to a significant slowdown in the computational model, the author will use a smaller number of random samples taken from the training dataset (called "batch size") in each epoch to bring into the model is 8. After training the model with the training dataset, the author proceed the test on the testing set to evaluate performance of the model.

3.2.4 Evaluating and comparing model performance at specific cases

In order to evaluate performance of VN-Index market index trend forecasting and have a basis for comparing the effectiveness in specific cases with different time periods, the study applies three commonly used evaluation metrics for classification problems as Accuracy, Precision and Recall (Devi & Radhika, 2018). In which, accuracy is the percentage of

correctly predicted observations out of the total number of observations in the testing dataset. Precision is the percentage of True Positive observations out of those observations are classified as positive. Recall is the percentage of True Positive observations out of those observations are actually Positive. The three indexes are calculated according to the formulas 12, 13, 14 respectively:

$$Accuracy = \frac{TP+TN}{n} \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

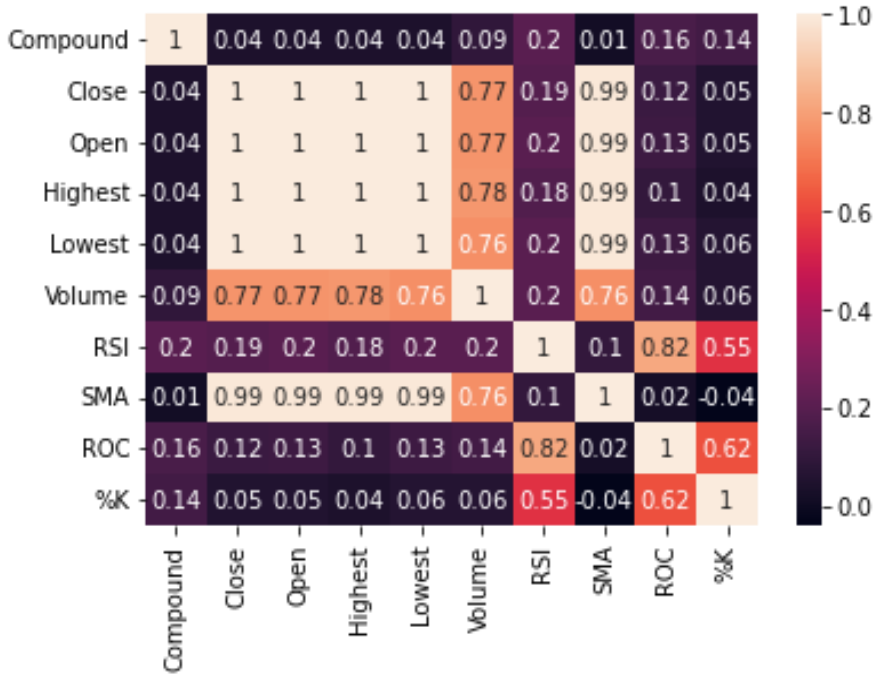
$$Recall = \frac{TP}{TP+FN} \quad (14)$$

From the above evaluation indexes, the author compares the effectiveness of the models with three specific cases. The first is with different forecasting time periods (1 day, 3 days, 9 days), which model has the highest accuracy and the forecast has the highest accuracy with which time period. Secondly, with the most effective forecasting time period, what is the precision and recall for each class. Thirdly, check to see if separating the dataset into two separate sets for forecasting and combining data for forecasting, will the results and influence levels be different?

4. RESEARCH RESULTS

4.1 Correlation analysis and descriptive statistics

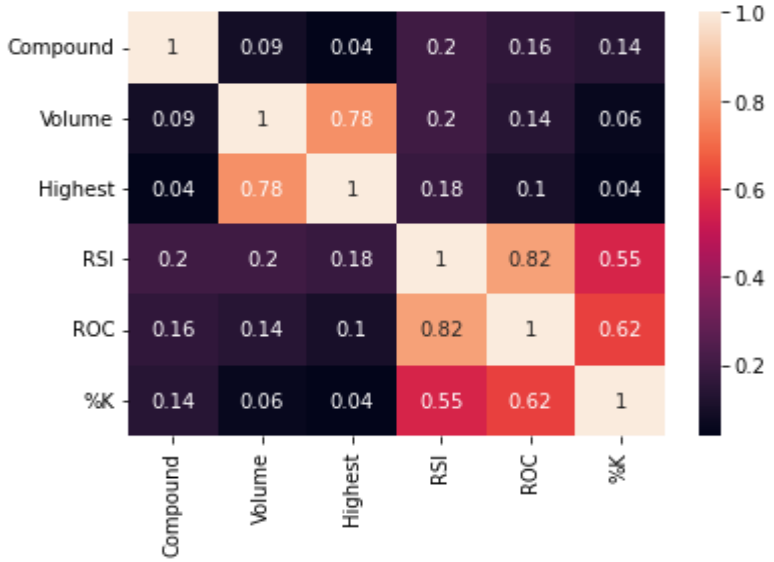
Figure 4. The correlation matrix between the independent variables



Source: The author

Figure 4 reflects the correlation of variables through the correlation coefficient. The analysis results show that among four variables including Close, Open, Highest, Lowest and the SMA indicator, there is a very high correlation (r = 0.99-1). According to Pallant (2007): "Multicollinearity exists when the independent variables are highly correlated (r = 0.9 or more)." Therefore, the author will delete these variables because the phenomenon of multicollinearity and correlation is too large. Large collinearity will affect the efficiency of the model, so it should be removed before being brought into the model.

Figure 5. Correlation matrix after removing highly correlated variables



Source: The author

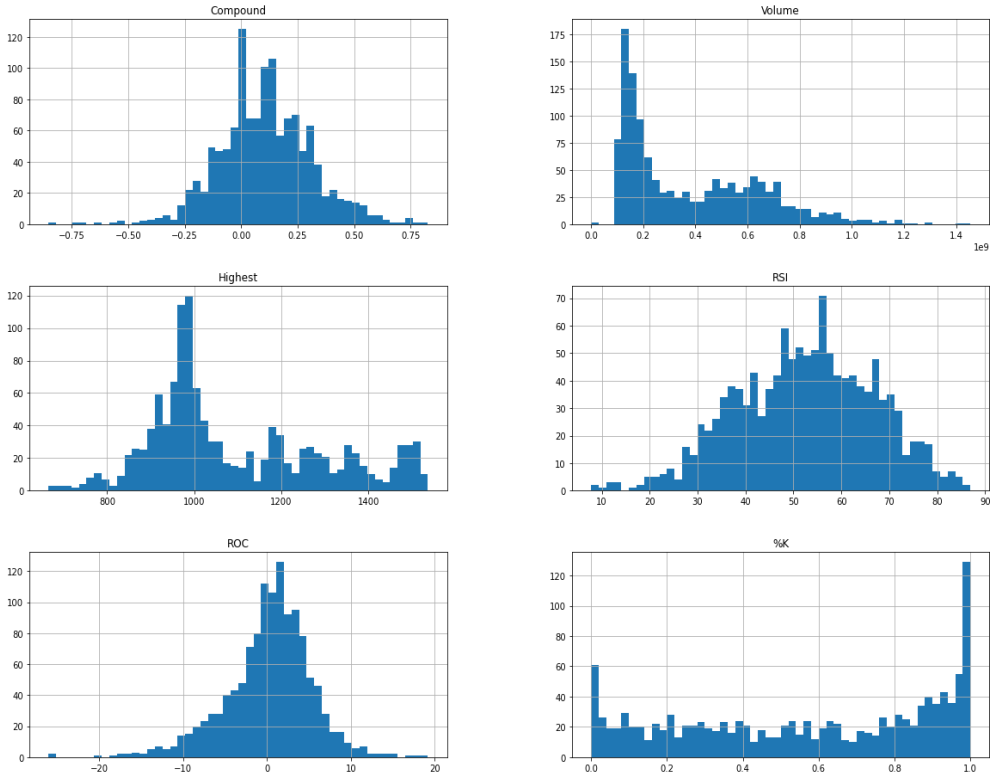
After removing the pairs of variables with high correlation that can cause interference and affect the efficiency of the model, the correlation matrix of the independent variables is at an acceptable level. In which, the pair of variables with the lowest correlation is between the Highest variable and the Compound variable with r = 0.04. The pair of variables with the highest correlation is ROC and RSI with r = 0.82. Part of the reason for this high correlation may be that the calculation methods all take the same value from the Closing value. Since r < 0.9, the author still keeps this pair of variables. The remaining variables have correlations ranging from 0.09 to 0.62.

Table 2. Descriptive statistics of independent variables

	Compound	**Volume**	**Highest**	**RSI**	**ROC**	**%K**
Count	1246	1246	1246	1246	1246	1246
Mean	0.106222	3.827874e+08	1093.25	52.523	0.208	0.567
Std	0.204702	2.632884e+08	199.275	14.346	5.139	0.335
Min	-0.851900	0.000000e+00	665.56	7.777	-26.051	0
25%	-0.012900	1.560150e+08	958.698	42.121	-2.217	0.263
50%	0.101728	2.846181e+08	1011.39	53.145	0.761	0.603
75%	0.232408	5.854058e+08	1246.28	63.066	3.43	0.892
Max	0.827100	1.454745e+09	1536.45	86.924	19.18	1

Source: Calculation of the author

Figure 6. Distribution chart of independent variables



Source: The author

Table 2 provides descriptive statistics about the independent variables and Figure 6 shows the distribution chart of the independent variables including: Compound, Volume, Highest and three technical indicators as RSI, ROC, %K. These are the variables that will be used as input variables for Machine Learning and Deep Learning models. For the Compound variable, the fluctuation range is mainly from -0.25 to 0.5, which focuses a lot at point 0 with 120 observations showing that many articles are neutral. Compound's distribution is highly concentrated in the range [0, 0.25] with an average score of 0.106 indicating that the news is more positive than the negative news. The trading volume is mainly distributed in the range [150000000, 500000000] with an average of 400000000 transactions per day. The variable Highest concentrates many observations at the peak of 1000 points. The highest score reached 1536.45 and the lowest score of the Highest variable reached 665.56 points. This is the variable with the highest volatility distribution with a standard deviation of 199. Meanwhile, the RSI variable mainly fluctuates in the range of 30 to 70, showing a regular uptrend and the occurrence of oversold/overbought is not much. The ROC indicator shows that the operating range is mainly in the range from -10 to 10. The Stochastic indicator with the %K variable, does not see a clear distribution, the two endpoints 0 and 1 occupy the most observations.

Table 3. Statistics of the number of observations in each class with time periods t

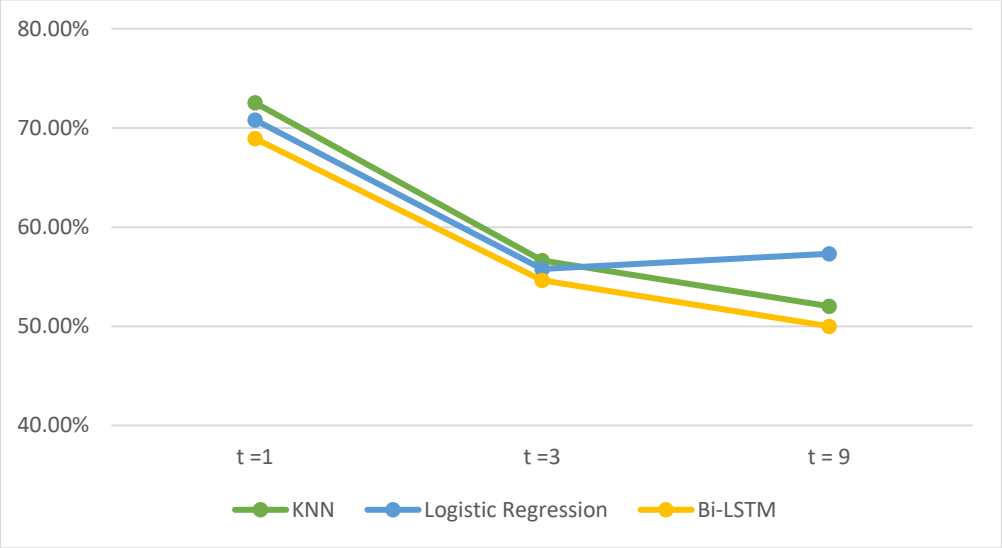
	t = 1	t = 3	t = 9
Uptrend (1)	811	724	712
Downtrend (0)	435	522	534

Source: Statistics of the author

With each different forecast time period, the number of observations in each class is different. In the case of t = 1 (predicting the trend for the next day), the ratio of the two classes is equivalent to 2:1. Meanwhile, the case with t = 3 is predicting the trend for the next 3 days and the case t = 9 is predicting the trend for the next 9 days, the two-class ratio is equivalent to 1.5:1.

4.2 Model results

Figure 7. Performance of the three models at different forecast time periods



Source: The author

In this section, the author evaluates the impact of the article headlines on the VN-Index in different time periods, by changing the value of t corresponding to the time periods (1 day, 3 days and 9 days). Figure 7 shows the accuracy of three models KNN, Logistic Regression and Bi-LSTM at different time periods. There are two outstanding points to be drawn from the above results:

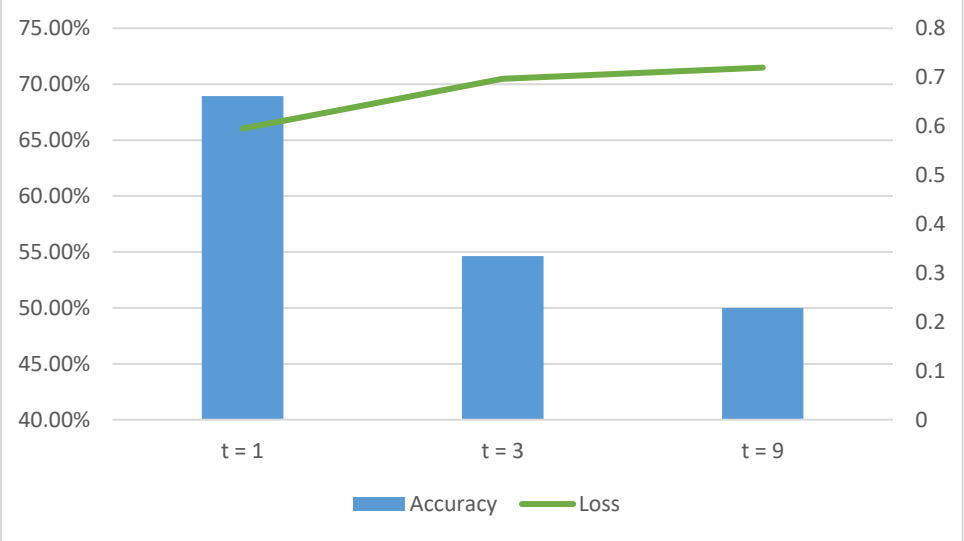
1) Of the three time periods, the trend forecasting for 24 hours after the publishing news has the highest accuracy. The performance of the forecasting models decreases when the forecasting day gets longer, proving that the hypothesis posed in the study by Minh Dang & Duc Duong (2016) is valid. In particular, the two authors believe that the article immediately affects the trading actions

of investors (buying, selling) in a short period of time (<=24 hours) and when the forecast period is longer, the news has less impact on investors' decisions.

2) Of the three models used for forecasting, KNN has the highest accuracy (72.54%) with k = 15, Logistic Regression has a lower accuracy of 70.81% with threshold = 0.564019 and Bi-LSTM achieves an accuracy of 68.93% with epoch = 30, batch-size = 8. In general, there is not much difference in the performance of the forecasting models.

Compared with Le Hong Hanh's model (2022) (with the highest accuracy of 60.1% with SVM) when forecasting the trend of VN-Index, it shows that in this research, all three models have higher accuracy when forecasting for the next day.

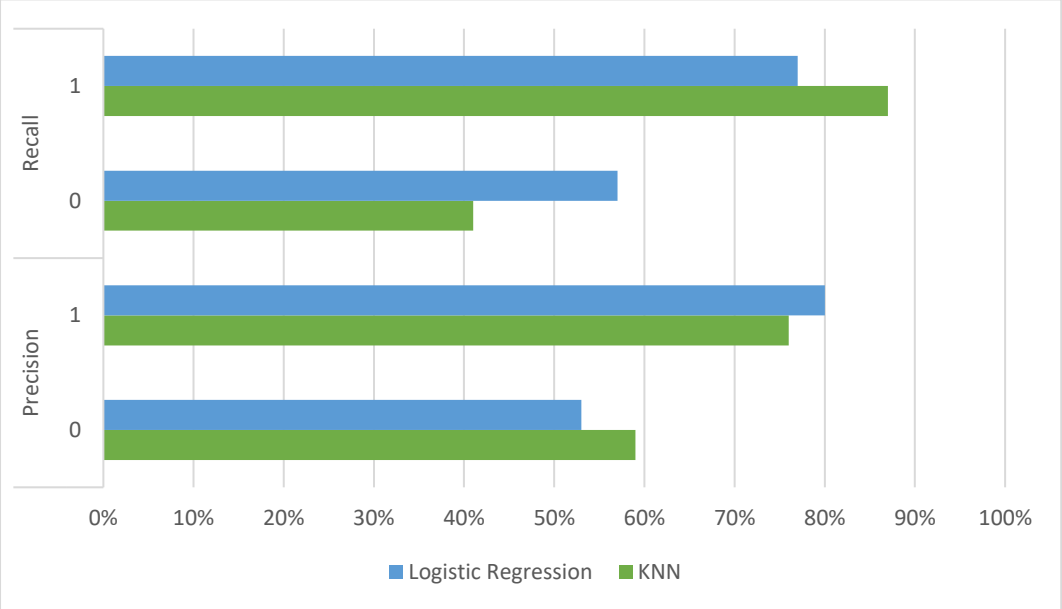
Figure 8. Performance of Bi-LSTM model at different time periods



Source: The author

Because in order to better evaluate the performance of the Deep Learning model as Bi-LSTM in forecasting the future trend of the VN-Index, the author conducts a test of the loss function and accuracy. At the 24-hour period, the model's forecasting accuracy achieved by 68.93%, but the loss function is still quite high (0.5949). The longer the forecasting time period is, the lower the accuracy is and the higher the loss function is. Although the Bi-LSTM model does not have higher accuracy than the above two Machine Learning models, when compared with other studies, the model still achieves an acceptable result.

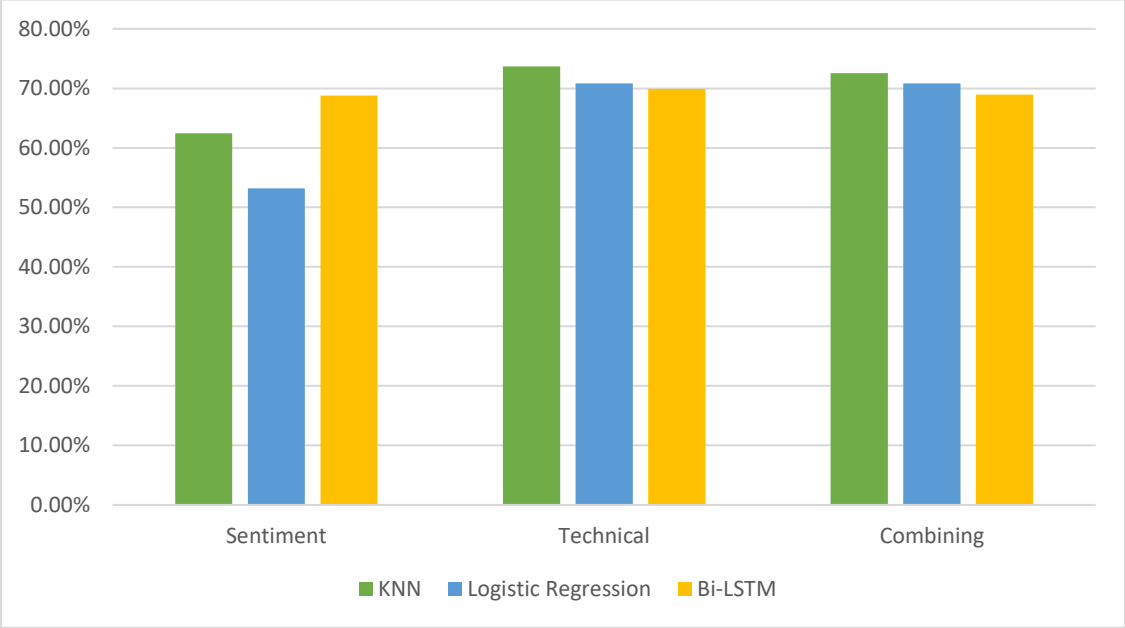
Figure 9. The detailed performance of Logistic Regression and KNN when forecasting for the next day



Source: The author

Instead of evaluating based on the loss function and accuracy like Bi-LSTM, the author uses two indexes including Precision and Recall to compare more clearly for each class of the output variable with Logistic Regression and KNN. Figure 9 shows that the KNN model has higher accuracy than Logistic Regression, but it does not mean that it is also better when forecasting for each class. Between the two labels Uptrend (1) and Downtrend (2), the Uptrend class is forecasted with higher accuracy than the Downtrend class. Part of the reason may be from the higher number of labels of the Uptrend class compared to the Downtrend class. In this study, the author wants to focus on the accuracy of the categorical forecasting model, so the result will pay more attention to Precision. In which, the rate of correctly forecasting Uptrend class of Logistic Regression (80%) is higher than that of KNN (76%). In which, with the label Downtrend, KNN (59%) is higher than Logistic Regression (53%). In general, it can be seen that there is little trade-off between Precision and Recall of the two models.

Figure 10. Performance comparison of three models with three data sets



Source: The author

In order to clarify the influence of the two input data sets of the forecasting model, the study has separated into three input data sources, including sentiment analysis dataset, technical analysis dataset and combining datase to compare as well as evaluate the performance clearly. Figure 10 shows that if only sentiment scores are used, Deep Learning model - Bi-LSTM has the highest accuracy with 68.79% - almost the same accuracy when combined with the technical indicator dataset. This shows that the Bi-LSTM model is really effective when used for forecasting in sentiment problems. Meanwhile, Logistic Regression has quite low accuracy compared to the other two models (53.18%). This can be seen as a limitation of Machine Learning models when compared to Deep Learning model in the field of sentiment analysis. With the input data as technical indicators, there is not too much difference between the three models. The case of combining two data sets also does not have a very clear difference for accuracy index.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

With efforts, a spirit of learning and serious work, I have almost completed the goals set out from the beginning and solved the research questions:

1) After finding out, researching concepts and theories related to the topic, the author used three main theories as the basis for predicting the trend of VN-Index including Efficient market theory, Behavioral finance theory and Information asymmetry.
2) Through the process of reviewing research and systematizing strengths and weaknesses, the author decided to use two libraries Vader and nltk to analyze sentiment on the article headlines. Besides, four technical indicators are used as independent variables including SMA, RSI, ROC and Stochastic.
3) With three forecasting time periods, t = 1 achieved the highest accuracy with the KNN model (72.54%). As t is larger, the accuracy of the model will decreases. Part of the reason could be due to the limitations of the data and the quality of the headlines as these news titles often have an impact over a fairly short period of time. Article titles have an effect on the stock market index and make it volatile. In which, the Bi-LSTM model when using the sentiment score of the news headline to make the input value achieved the highest accuracy with 68.79%.
4) From the results of three models including Logistics Regression, KNN and Bi-LSTM, the study compared the average performance of each model through evaluating metrics such as Accuracy, Precision and Recall. The study achieved an accuracy of over 65% - this is a fairly objective result which is good enough for a stock market index forecast model. Depending on each dataset, the models have different performance. Typically, if using sentiment scores, Deep Learning model has higher performance. While if using technical indicators and combined data sets, two Machine Learning models have higher accuracy. In addition, the analysis and evaluation of sentiment on article headlines and the use of technical indicators as input features shows a new approach to be able to combine traditional methods and Machine Learning/Deep Learning methods. From there, investors can apply to get more references about the volatility trend of the market index to have more consulting sources to help make investment decisions or not.

Although the research has been carried out and received positive results as expected by the author, this study still has some limitations. Firstly, the dataset used in this study is only taken from one newspaper page, so it does not show diversity. The second is the limitation in data extraction and conversion to sentiment score. Third, the input factors are quite few when using only three technical indicators and the time of data is limited, which can affect the performance of the model. Finally, the topic is currently only effective when forecasting for the next day and limited for the following forecast days.

5.2 Recommendations

From the limitations of the study mentioned above, the author proposes some new solutions and recommendations. The projects in the future may need to crawl and synthesize more online newspapers so that the data source for analyzing sentiment will be more diverse. Specifically, the text mining in reputable newspapers specializing in finance - securities so that when scoring sentiment will be more objective than a page source. On the other hand, the following research should exploit data with a longer time period so that the model will learn and be trained more effectively. In the upcoming research papers, the author will develop this research paper more when it can be applied to enterprises listed on the stock exchange to forecast trends, then split buy/sell/hold investment strategy according to risk appetite to help come up with better models with better results to help investors have more information to make decision.

REFERENCES

1) Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. The quarterly journal of economics, 84(3), 488-500.
2) Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016, June). Deep learning for stock prediction using numerical and textual information. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.
3) Almahirah, M. S., VNS, M. J., Sharma, S., & Kumar, S. (2021). Role of market microstructure in maintaining economic development. Empirical Economics Letters, 20(2.2021).
4) Ang, J. S., & Pohlman, R. A. (1978). A note on the price behavior of Far Eastern stocks. Journal of International Business Studies, 103-107.
5) Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. Expert Systems with applications, 36(3), 5932-5941.
6) Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, 7(4), 190-207.
7) Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 450-453).
8) Bujari, A., Furini, M., & Laina, N. (2017, January). On using cashtags to predict companies stock trends. In 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC) (pp. 25-28). IEEE.
9) Chae, J. (2005). Trading volume, information asymmetry, and timing information. The journal of finance, 60(1), 413-442.
10) Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).
11) Devi, S. S., & Radhika, Y. (2018). A survey on machine learning and statistical techniques in bankruptcy prediction. International Journal of Machine Learning and Computing, 8(2), 133-139.
12) Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. The Annals of Statistics, 1310-1319.
13) Devroye, L. (1988). Automatic pattern recognition: A study of the probability of error. IEEE Transactions on pattern analysis and machine intelligence, 10(4), 530-543.
14) Devroye, L., & Lugosi, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. The Annals of Statistics, 2626-2637.

15) Di Persio, L., & Honchar, O. (2017). Recurrent neural networks approach to the financial forecast of Google assets. International journal of Mathematics and Computers in simulation, 11, 7-13.
16) Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.
17) Faria, J. A. D., & Gomes, S. M. D. S. (2013). The effects of information asymmetry on budget slack: An experimental research.
18) Garman, M. B. (1976). Market microstructure. Journal of financial Economics, 3(3), 257-275.
19) Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.
20) Glosten, L. R., & Harris, L. E. (1988). Estimating the components of the bid/ask spread. Journal of financial Economics, 21(1), 123-142.
21) Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of financial economics, 14(1), 71-100.
22) Gorgulho, A., Neves, R., & Horta, N. (2011). Applying a GA kernel on optimizing technical analysis rules for stock picking and portfolio composition. Expert systems with Applications, 38(11), 14072-14085.
23) Hiển, N. Đ. (2013). Hành vi của nhà đầu tư trên thị trường chứng khoán Việt Nam.
24) Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.
25) Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
26) Huang, R. D., Masulis, R. W., & Stoll, H. R. (1996). Energy shocks and financial markets. Journal of Futures markets, 16(1), 1-27.
27) Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications, 3(5), 605-610.
28) Kahneman, D., & Tversky, A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen.
29) Kingma, D. P., Mohamed, S., Jimenez Rezende, D., & Welling, M. (2014). Semi-supervised learning with deep generative models. Advances in neural information processing systems, 27.
30) Lamoureux, E. L., Pallant, J. F., Pesudovs, K., Rees, G., Hassell, J. B., & Keeffe, J. E. (2007). The impact of vision impairment questionnaire: an assessment of its domain

structure using confirmatory factor analysis and Rasch analysis. Investigative ophthalmology & visual science, 48(3), 1001-1006.

31) Le Hong, H., Nguyen, N. N., Nguyen, T. L., Nguyen, L. D., & Nguyen, N. H. (2022). Stock Market Prediction: The Application of Text-Mining in Vietnam. VNU JOURNAL OF ECONOMICS AND BUSINESS, 2(2).

32) Li, Y., Zheng, Z., & Dai, H. N. (2020). Enhancing bitcoin price fluctuation prediction using attentive LSTM and embedding network. Applied Sciences, 10(14), 4872.

33) Liu, Q., & Tse, Y. (2017). Overnight returns of stock indexes: Evidence from ETFs and futures. International Review of Economics & Finance, 48, 440-451.

34) Madhavan, A. (2000). Market microstructure: A survey. Journal of financial markets, 3(3), 205-258.

35) Madhavan, A., Richardson, M., & Roomans, M. (1997). Why do security prices change? A transaction-level analysis of NYSE stocks. The Review of Financial Studies, 10(4), 1035-1064.

36) Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. Ieee Access, 6, 55392-55404.

37) Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W., & Wills, G. (2020). A survey on machine learning for stock price prediction: Algorithms and techniques.

38) O'Hara, F. (1995). The collected poems of Frank O'Hara. Univ of California Press.

39) Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, 109-131.

40) Omoruyi, A., & Osad, O. I. (2018). A test of market microstructure model: Evidence from Nigerian Stock Market. Amity Journal of Finance, 3 (1), 1-13.

41) Ramdeen, C., Santos, J., & Chatfield, H. K. (2007). An examination of impact of budgetary participation, budget emphasis, and information asymmetry on budgetary slack in the Hotel Industry. Article. William F. Harrah of Collage of Hotel Administration University of Nevada, Las Vegas.

42) Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: three types of fakes. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.

43) Selden, G. C. (1912). Psychology of the stock market. Ticker.

44) Shiller, R. C. (2000). Irrational exuberance. Philosophy and Public Policy Quarterly, 20(1), 18-23.

45) Sunny, M. A. I., Maswood, M. M. S., & Alharbi, A. G. (2020, October). Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. In 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES) (pp. 87-92). IEEE.

46) Thaler, R. H. (1985). Illusions and mirages in public policy. In Environmental Impact Assessment, Technology Assessment, and Risk Analysis: Contributions from the Psychological and Decision Sciences (pp. 567-581). Springer Berlin Heidelberg.
47) Thiều, L. T. L. (2006). Thông tin bất cân xứng & các quyết định tài chính của các công ty cổ phần Việt Nam (Doctoral dissertation, Đại học Kinh tế TP Hồ Chí Minh).
48) Thủy, P. B. G., Phúc, N. T., & Trọng, N. V. ĐẶC ĐIỂM HỘI ĐỒNG QUẢN TRỊ VÀ THÔNG TIN BẤT CÂN XỨNG: ẢNH HƯỞNG ĐIỀU TIẾT CỦA LOẠI HÌNH DOANH NGHIỆP.
49) Upadhyay, A., Bandyopadhyay, G., & Dutta, A. (2012). Forecasting stock performance in indian market using multinomial logistic regression. Journal of Business Studies Quarterly, 3(3), 16.
50) Venkatesh, P. C., & Chiang, R. (1986). Information asymmetry and the dealer's bid-ask spread: A case study of earnings and dividend announcements. The Journal of Finance, 41(5), 1089-1102.
51) Wang, F., Shieh, S. J., Havlin, S., & Stanley, H. E. (2009). Statistical analysis of the overnight and daytime return. Physical Review E, 79(5), 056109.
52) Wenjuan, W. (2017). Study on the Duration of Market Microstructure Theory. International Journal of Business and Management, 12(10).
53) Yang, M., Moon, J., Yang, S., Oh, H., Lee, S., Kim, Y., & Jeong, J. (2022). Design and implementation of an explainable bidirectional lstm model based on transition system approach for cooperative ai-workers. Applied Sciences, 12(13), 6390.
54) Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Third IEEE international conference on data mining (pp. 427-434). IEEE.

APPENDIX

```
#Load the libraries
import pandas as pd
from datetime import datetime
import numpy as np
import ta
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score,
roc_auc_score, recall_score, roc_curve
from sklearn.model_selection import train_test_split, KFold,
cross_val_score,GridSearchCV,StratifiedKFold
from numpy import sqrt, argmax
from matplotlib import pyplot
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.optimizers import Adam
from tensorflow.keras import layers
from keras.callbacks import EarlyStopping
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, SimpleRNN, Dense, LSTM, Bidirectional

#Calculate technical indicators
data = pd.read_csv('Data_Vnindex_demo.csv')
data.Date = pd.to_datetime(data.Date)
data['RSI'] = ta.momentum.rsi(data['Close'])
data['SMA'] = ta.trend.sma_indicator(data['Close'], 10)
data['ROC'] = ta.momentum.roc(data['Close'])
data['%K'] = ta.momentum.stochrsi_k(data['Close'])
data = data.dropna()
data.to_csv('Data_VNINDEX.csv')

#Prepare the dataset
```

```
df.Date = pd.to_datetime(df.Date)
vader = SentimentIntensityAnalyzer()
scores = df['Title'].apply(vader.polarity_scores).tolist()
df_scores = pd.DataFrame(scores)
df = df.join(df_scores, rsuffix='_right')
df = df.rename(columns={'compound': 'Compound'})
col = ['Date', 'Compound']
df = df[col]
df = df.resample('d', on='Date').mean().dropna(how='all')
df = df.sort_values(by='Date', ascending=True)
df1 = pd.read_csv('Data_VNINDEX.csv')
df1.Date = pd.to_datetime(df1.Date)
data = df.merge(df1, how='right', on='Date')
#Data with t = 1
data['Open'] = data['Open'].shift(-1)
data = data.dropna()
data['Trend_t1'] = data['Open']-data['Close']
data.loc[data['Trend_t1'] >= 0, 'Trend_t1'] = 1
data.loc[data['Trend_t1'] < 0, 'Trend_t1'] = 0
data.info()
data.reset_index()
col_all = ['Date', 'Compound', 'Close', 'Open', 'Highest', 'Lowest', 'Volume', 'MACD',
'Signal', 'RSI', 'SMA', 'ROC', 'BollingerBand', '%K', 'Trend_t1']
data = data[col_all]
explanatory_feature = ['Compound', 'Volume', 'Highest', 'RSI', 'ROC', '%K']
target = ['Trend_t1']
# Correlation matrix
matrix = data[explanatory_feature].corr()
# Plot correlation matrix
sns.heatmap(matrix.round(2), annot=True)
plt.show()
data[explanatory_feature].describe()
%matplotlib inline
#Creating histogram for numerical attributes
data[explanatory_feature].hist(bins=50, figsize=(20,15))
plt.show()
# Count the number of each class of the target variable
data['Trend_t1'].value_counts()
```

```
X = data[explanatory_feature].values
y = data[target].values
X_train, y_train = X[:900], y[:900]
# X3_val, y3_val = X3[60000:65000], y3[60000:65000]
X_test, y_test = X[900:], y[900:]
X_train.shape, X_test.shape, y_train.shape, y_test.shape
scalar = MinMaxScaler()
scalar.fit(X_train)
X_train = scalar.transform(X_train)
scalar.fit(X_test)
X_test = scalar.transform(X_test)

# Fit a model Logistic regression
model = LogisticRegression(random_state = 0)
model.fit(X_train, y_train)
# Predict probabilities
y_pred = model.predict(X_test)
# Confusion matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
print(classification_report(y_test,y_pred))
print('Logistic accuracy: ', accuracy_score(y_test, y_pred))
# Keep probabilities for the positive outcome only
y_pred = model.predict_proba(X_test)
y_pred = y_pred1[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
# Calculate the g-mean for each threshold
gmeans = sqrt(tpr * (1-fpr))
# Locate the index of the largest g-mean
ix = argmax(gmeans)
print('Best Threshold=%f, G-Mean=%.3f' % (thresholds[ix], gmeans[ix]))
# plot the roc curve for the model
pyplot.plot([0,1], [0,1], linestyle='--', label='No Skill')
pyplot.plot(fpr, tpr, marker='.', label='Logistic')
pyplot.scatter(fpr[ix], tpr[ix], marker='o', color='black', label='Best')
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
pyplot.legend()
```

```
pyplot.show()
# Test with new threshold
threshold = 0.564019
y_pred2 = (model.predict_proba(X_test)[:, 1] > threshold).astype('float')
accuracy = accuracy_score(y_test, y_pred2)
cm1 = confusion_matrix(y_test, y_pred2)
print(cm1)
print(classification_report(y_test,y_pred2))
print("Logistic Accuracy: %.2f%%" % (accuracy * 100.0))

# Fit a model with default n_neighbors = 5
model1 = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
model1.fit(X_train, y_train)
# Predict probabilities
y_pred1 = model1.predict(X_test)
# Report result of test data following model
cm = confusion_matrix(y_test, y_pred1)
accuracy = accuracy_score(y_test, y_pred1)
print(cm)
print(classification_report(y_test,y_pred1))
print("KNN Accuracy: %.2f%%" % (accuracy * 100.0))
acc = []
for i in range(1,40):
   neigh = KNeighborsClassifier(n_neighbors = i).fit(X_train,y_train)
   yhat = neigh.predict(X_test)
   acc.append(metrics.accuracy_score(y_test, yhat))
plt.figure(figsize=(10,6))
plt.plot(range(1,40),acc,color = 'blue',linestyle='dashed',
        marker='o',markerfacecolor='red', markersize=10)
plt.title('accuracy vs. K Value')
plt.xlabel('K')
plt.ylabel('Accuracy')
print('Maximum accuracy:-',max(acc),'at K =',acc.index(max(acc)))
# Fit a model with default n_neighbors = 15
model1 = KNeighborsClassifier(n_neighbors = 15, metric = 'minkowski', p = 2)
model1.fit(X_train, y_train)
# Predict probabilities
y_pred1 = model1.predict(X_test)
```

```
# Report result of test data following model
cm = confusion_matrix(y_test, y_pred1)
accuracy = accuracy_score(y_test, y_pred1)
print(cm)
print(classification_report(y_test,y_pred1))
print("KNN Accuracy: %.2f%%" % (accuracy * 100.0))

#Fit model with Bi-LSTM
model2 = Sequential([layers.Input((6, 1)),
                layers.Bidirectional(LSTM(128, return_sequences=True)),
                layers.Dense(256, activation='relu'),
                layers.Dropout(0.2),
                layers.Dense(256, activation='relu'),
                layers.Dropout(0.2),
                layers.Dense(1, activation='sigmoid')])
model2.compile(loss='binary_crossentropy',
            optimizer='adam',
            metrics=['acc'])
callback = EarlyStopping(monitor='loss',patience = 10)
history1=model2.fit(X_train,y_train, callbacks=[callback], epochs=500, batch_size=8)
# Evaluate the keras model
model2.evaluate(X_train,y_train)
# Summarize history for loss
plt.plot(history1.history['loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train'], loc='upper left')
plt.show()
test_result=model2.evaluate(X_test,y_test)
print(test_result)
```