IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2024, accepted 5 August 2024, date of publication 12 August 2024, date of current version 21 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441755

RESEARCH ARTICLE

Study on the Qualitative Cohesion in Bitcoin Market Price Prediction

NAMJAE CHO, JAE HYUN BYUN, AND GISEOB YU
School of Business, Hanyang University, Seoul 04763, South Korea
Corresponding author: Giseob Yu (yugs@hanyang.ac.kr)

ABSTRACT Over time, various methodologies have been introduced for predicting the cryptocurrency market. While numerous studies have explored different variables, research incorporating the actual sentiments of investors has been scarce. In this study, we aimed to improve cryptocurrency market predictions by considering the qualitative cohesion. We built upon the existing LSTM model and extended our analysis to include RoBERTa and DistilBERT models through text mining. The results revealed that RoBERTa and DistilBERT incorporating investor sentiment outperformed the LSTM model in terms of prediction accuracy. Notably, the DistilBERT model, known for its exceptional word and context analysis, demonstrated the highest predictive power, followed by RoBERTa and the LSTM model. These findings underscore the importance of directly analyzing investor psychology in future market analyses. Furthermore, focusing on both individual words and contextual meaning is expected to yield even better market prediction results.

INDEX TERMS Cohesion, cryptocurrency, DistilBERT, RoBERTa, LSTM.

I. INTRODUCTION

The rapid growth of the cryptocurrency market over the past decade has attracted a variety of investors to the market [1], and various studies have been conducted to predict the future of the market [2]. Generally, market prediction research methodologies can be classified into fundamental analysis, which predicts by analyzing information such as financial factors and competitive factors, and technical analysis [3]. However, with the recent advancement of computing power [4], research applying artificial intelligence technologies is being conducted for improving market predictability.

For the prediction of closing prices, research has mainly been conducted using the LSTM (Long Short Term Memory) model. LSTM has advantages such as speed improvement and prediction performance improvement compared to existing models [5]. In addition, LSTM is a representative time series analysis model that can solve the long-term dependency problem that occurs in existing analysis models due to the Vanishing Gradient problem [6], [7].

While there have been many attempts to improve the accuracy of cryptocurrency market predictions using various deep learning analysis models, there have been problems such as sharp fluctuations in the market and a decrease in prediction rates when certain patterns occur [8]. To solve these problems, research that finds and applies new variables [7], and predictive research using cohesion [9] have been conducted.

The reason for employing various methods to forecast the cryptocurrency market stems from the high volatility and complexity inherent in this market. If accurate predictions of the cryptocurrency market can be achieved, it would aid investors in enhancing their investment strategies and risk management, thereby facilitating more stable investment management. However, previous studies have limitations such as confirming new variables or approaching the quantitative aspect of cohesion. In this study, we aim to overcome these limitations by collecting and analyzing qualitative elements of data. We aim to enhance the market prediction model that considers qualitative aspects using the Robustly optimized BERT approach (RoBERTa) sentiment model and DistilBERT. For this, we plan to compare and analyze the results of the cohesion model and the qualitative cohesion model set in this study.

The significance of this study is as follows. First, we presented the necessity and importance of qualitative cohesion research and analysis. It broadens the academic perspective

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

on qualitative cohesion, and by directly collecting and analyzing text reflecting the psychology of investors, it shows high result accuracy. Second, based on the results of this study in the cryptocurrency market, it is expected that it can be utilized in various financial markets such as the stock market and investment product prediction through the analysis model reflecting the psychology of future investors.

Following the introduction in Chapter 1, the composition of this study is as follows. In Chapter 2, we will learn about Cohesion, LSTM, RoBERTa, and DistilBERT, and in Chapter 3, we will explain the research methodology. In Chapter 4, we describe the prediction results of the analysis model, and in Chapter 5, we look at the conclusion and future research.

II. THEORETICAL BACKGROUND

A. COHESION IN SOCIAL NETWORKS

In Social Network Analysis (SNA), cohesion is a fundamental metric that is essential for comprehending the dynamics of social networks. Cohesion is applied across variety of disciplines within SNA. Cohesion is a tool used to examine user interactions on social media platforms like Twitter. Researchers can determine how quickly information spreads, measure user influence, and monitor interaction patterns by looking at the connection strength of posts [10]. Understanding the flow of its information and impact is critical in domains like public relations, political campaigns, and marketing, where this kind of study is vital. In SNA, cohesion is a complex metrix that offers important insights into the type and quality of interactions in diverse contexts. It is not merely a measure of network density or connection strength.

Previous research indicates that a network with many connections and strong cohesion is more likely to have access to a variety of sources [11] and that information spreads more quickly [12]. However, when cohesion increases, information circulation decreases [13]. Additionally, because information circulates inside a network with a high degree of similarity, there are limits on the diversity of information available [14]. For example, the study [9] investigates the relationship between social media discussions and the Bitcoin market, focusing on how network cohesion within these discussions influences market predictions. The study analyze whether conversations on platforms like Twitter can provide valuable signals for forecasting Bitcoin prices or whether they merely contribute noise. The study find that when social media discussions are more cohesive—meaning participants are closely connected and engage in focused conversations—the information shared tends to be more predictive of Bitcoin market trends. Conversely, when discussions are fragmented or less cohesive, the ability to predict market movements diminishes, suggesting that the structure of online networks plays a crucial role in distinguishing between meaningful signals and irrelevant noise in social media data. This study has some strengths, such as its innovative focus on network cohesion and its interdisciplinary relevance. On the other hand, it has a limited scope and understates aspects of social media, such as sentiment analysis

According to recent research, relying solely on quantitative factors for cohesion analysis may lead to unintended outcomes. However, the quality of cohesion, which is essential in various contexts, including financial markets, can significantly benefit from incorporating perspectives such as investor sentiment. This approach is exemplified in a study demonstrating how sentiments expressed in online comments impact the predictability of cryptocurrency market movements. By integrating sentiment analysis into cohesion evaluations, researchers can achieve a more comprehensive understanding of market dynamics, highlighting the multifaceted nature of investor behavior and market predictability.

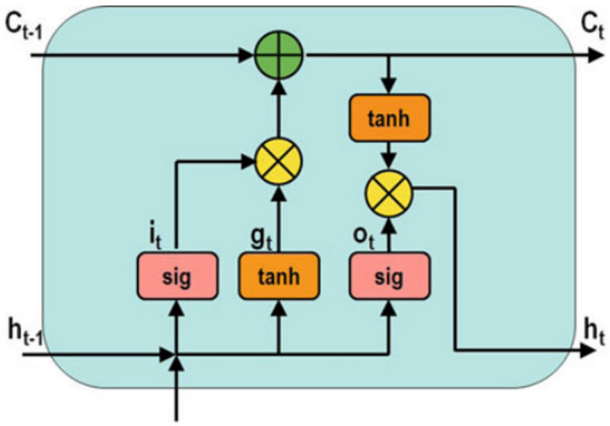

FIGURE 1. An initial LSTM cell model [33].

B. ORIGINAL LONG SHORT TERM MEMORY (LSTM)

The Recurrent Neural Network (RNN), which is mainly used for future prediction by analyzing trend fluctuations and irregular fluctuation data, is a model suitable for analyzing variable-sized data [15]. However, because RNN uses all collected data due to the nature of time series analysis, it requires a long learning time and sometimes loses some information [5].

The model developed to solve the problems of RNN is the LSTM model, which is designed to predict the future by considering macroscopic analysis data [5]. LSTM is an improved model to overcome the limitations of RNN, and in particular, it has overcome the problem of Vanishing Gradient, where the accuracy drops because past data is not learned [6].

The data analysis step of the LSTM is a model that adds a cell state to the Hidden Layer on the RNN model. It consists of a Forget Gate that decides what information to delete from the Cell State, an Input Gate that decides what information to remember in the Cell State from the newly incoming information. Lastly, an Output Gate that chooses what information to output in the next step.

LSTM has been used for predictive analysis in various fields, such as human behavior, market prices, and demand.

In a study that analyzed the movement path based on the current location of pedestrians using LSTM [4], it predicted the movement path of individuals and groups with a high probability compared to RNN-based analysis.

In this study, researchers proposed a new analysis model called Social LSTM. In the results of the analysis applying the LSTM-based nonlinear model in the stock market where prediction is difficult, effective stock prediction results could be derived [16]. In a study to predict The S&P 500 index based on the Layer, market basic data, macroeconomic data, technical factors, etc. were used for analysis, and there is also a conclusion that the Single Layer LSTM is more effective than Mutiplelayer LSTM Models [17].

LSTM has also been used in various perspectives in cryptocurrency market analysis. Various perspectives and factors have been reflected in the research that has been conducted, such as research on strategic use in Bitcoin trading using LSTM analysis results [18], accuracy analysis on Bitcoin price prediction [19], and comparative analysis research on Bitcoin and gold price prediction using LSTM-based model [20]. As confirmed in previous studies, LSTM can be said to be a deep learning model optimized for time series data analysis that overcomes the disadvantages of RNN.

In previous studies, the development of market prediction models was primarily conducted through a quantitative approach to data. The state of qualitative approach research is inadequate, and this study aims to fill that gap. Considering the psychological part of investors in the cryptocurrency market using LSTM as the basic model, we focus on improving its accuracy by adding qualitative factors.

C. ROBERTA MODEL

As Social Media has become a platform that reflects individual sentiments in daily life [21], various natural language processing methods for analyzing consumer sentiment have become an important research field [22]. As one of the hybrid deep learning models, Transformation models were presented to address the issue of vocabulary's distance dependency between contexts [23]. The first language representation modeling introduced based on this model is Bidirectional Encoder Representations from Transformers (BERT) [24]. One of the BERT models, RoBERTa, is a model developed based on the Transformer family [25]. While RNN-based models are effective in sequence-to-sequence modeling analysis, RoBERTa is effective in sequence-to-vector modeling analysis.

The analysis process of RoBERTa generally uses a large amount of data to learn using batch size, and then removes the objective function of 'Next Sentence Prediction'. After that, it improves the performance of BERT by learning the model using a longer sequence and dynamically assigning 'Masking' [25]. The RoBERTa model has been fine-tuned based on about 124 million documents tweeted from 2018 to 2021, and it is expected to contribute to the improvement of future predictions of the cryptocurrency market that considers qualitative aspects by combining with the existing LSTM.

D. DISTILBERT MODEL

As the Large-scale Pre-Trained Language (LPL) model based on natural language has gradually developed [25], [26], some problems such as the increase in cost for analysis [27] and the need for high-performance computers [26] have occurred. To solve these problems, various analysis models have been developed. Among them, the DistilBERT model developed by HuggingFace [28] is one of the most effective and optimized models for the LPL model. The biggest feature of DistilBERT is that it distills information to solve the problems of the existing LPL [28]. The knowledge distillation method was implemented in its completed form through a generalization process after its introduction in 2006 [29], [30]. DistilBERT has the advantage of effectively operating even with low computer performance, to the extent that it can be implemented on mobile devices [28].

As results of using DistilBERT, it showed a 40% reduction in data size, more than 60% speed improvement, and a 97% level of natural language understanding ability compared to the existing BERT model [28]. It is an effective natural language analysis model that can reduce analysis costs as a small, fast, and lightweight analysis model compared to BERT.

In this study, we plan to use DistilBERT, which has been fine-tuned with the Stanford Sentiment Treebank (SST-2), a sentiment analysis dataset. SST-2 is specialized in qualitative analysis of language and context as it also assigns sentiment labels to each word and phrase in the sentence. We aim to apply the LSTM and DistilBERT models to the prediction of the cryptocurrency market affected by qualitative factors.

III. METHODS

A. HYPOTHESIS DEVELOPMENT

In social media, cohesion refers to a factor that influences the exchange of information on a network [9]. Networks with high cohesion have the advantages of efficient communication [31] and improved accuracy of information [32]. In other words, it means that communication between users can be improved on a network with high cohesion, and the circulation of information can proceed quickly.

In addition to the circulation of information, cohesion also affects user behavior. According to a cohesion-related study conducted in the field of network games [33], users on a social game network were directly influenced by nearby players in their willingness to pay within the game. Members of a group with high cohesion also actively participate in group decision-making [34]. As cohesion increases, the diversity and exchange of information decrease [13], and there can be a problem of circulating similar information within one network [14]. However, as mentioned earlier, high cohesion can have a greater impact on user decision-making, so high cohesion is a very important factor in market prediction.

In recent social network analysis related to cryptocurrency, networks with low cohesion had a greater impact on Bitcoin prices than those with high cohesion [9]. However, in the

study, the qualitative aspect was excluded in the analysis of network cohesion. To fill this gap, we conduct an analysis by adding qualitative factors along with the quantitative factors that have been used in the prediction of the existing cryptocurrency market. The hypotheses set for this are as follows.

H0: An LSTM (Long Short-Term Memory) model that incorporates qualitative aspects of cohesion as a variable will not demonstrate superior performance compared to an LSTM model that includes only traditional, quantitative measures of cohesion.

H1: An LSTM (Long Short-Term Memory) model that incorporates qualitative aspects of cohesion as a variable will outperform an LSTM model that includes only traditional, quantitative measures of cohesion.

B. VARIABLES IN THE ANALYSIS

Previous research has shown the effectiveness of LSTM models in forecasting Bitcoin values. This study examines the residuals of three LSTM models that included variables including the closing prices of gold and bitcoin, the Volatility Index (VIX) value, the S&P index, the trading volume of bitcoin, normal cohesion, integrated cohesion with RoBERTa and DistilBERT. In order to compute residuals, the first LSTM model is trained using the following variables: ''Gold closing price,[1]'' ''Bitcoin closing price,'' ''VIX value,'' ''S&P index,'' ''Bitcoin trading volume,'' and normal cohesion. The cohesion integrates with qualitative characteristics using RoBERTa is included in the second LSTM model, which eliminates normal cohesion and from which residuals are generated. The third LSTM model uses cohesiveness as a variable to calculate residuals, with qualitative aspects evaluates using DistilBERT.

C. DATA AND METHOD PROCESS

The data for the analysis was collected from Bitcointalk.com, a discussion website related to Bitcoin which is a representative coin in the cryptocurrency market. A total of 190,796 text data were collected from January 1, 2017 to September 17, 2023. The collected quantitative data are Bitcoin price, trading volume, volatility index (VIX), cohesion, and S&P price. The method of calculating cohesion used in previous research was here.

However, in this study, the following formula is added to measure qualitative cohesion. In the RoBERTa sentiment model, the sentiment text was quantified by multiplying -1 when the text data is negative, 0.1 when it is neutral, and 1 when it is positive.

RoBERTa

$$Score = C \times (-1 \times RobertaNeg + 0.1 \times RobertaNeu + 1 \times RobertaPos)$$

In the DistilBERT sentiment model, the weight of qualitative cohesion was reflected in the result value by multiplying -1 in case of negativity and 1 in case of positivity.

[1] ''Closing price'' refers to the price at the end of the trading day.

DistilBERT:

$$Score = C \times (-1 \times DistillNeg + 1 \times DistillPos)$$

The data preprocessing was conducted as follow. In order to extract time from text data and arrange it by year, month, day, hour, and minute, we first crawled the data, which included date, time, and post. Second, we counted both the total and the number of quotes. The groupby function was used to differentiate the text data on a daily basis. The daily amount of postings and the frequency with which the phrase ''Quote from'' appeared were computed. The Average-Degree technique was utilized in the research to calculate cohesion [9]. Thirdly, data that was empty or duplicated was eliminated. The text data's only characters and digits were kept. Next, we executed the tokenization process, which involved splitting the cleaned text into individual tokens, typically words or phrases, based on whitespace and punctuation marks. During this process, all tokens were normalized by converting them to lowercase to ensure consistency across the dataset.

Fourth, the maximum length for sentiment analysis with RoBERTa and DistilBERT is 510 characters. Consequently, the original date was added and the text was divided into 510 words. After text splitting, the original dataset's approximately 190,000 rows increased to about 330,000. After a daily average of the sentiment analysis results for these 330,000 entries, roughly 1,800 rows were produced. Fifthly, the closing prices for the S&P index, Bitcoin, gold, and VIX were displayed in a string format with commas inserted between the values. As a result, the commas were eliminated and the format changed to float.

In the Sixth, B, M, and K units were used to represent the volume of Bitcoin trading. 'K' multiplied numbers by 1,000, 'M' by a million, and 'B' by a billion. To lower the quantity of the data, compute the sentiment analysis results' daily average. Finally, the following variables were employed independently: ''gold closing price,'' ''Bitcoin closing price,'' ''VIX value,'' ''S&P index,'' ''Bitcoin trading volume,'' and ''daily average value of sentiment.''

The input layer of the LSTM model was composed of 200 units and 30% of the neurons were deactivated, and the hidden layer deactivated 20% of the 150 unit neurons. To reduce the problem of technical loss in neural networks, the activation function ReLU was applied. The optimization algorithm used Adam, and the loss function used MSE. The flow of this study is illustrated as follows (Figure 2).

For analyzing the cohesion, we apply the method suggested by [9]. The cohesion explains closeness and the connection strength within a network. Because of the characteristics in the network, the connection strength and closeness is crucial diffusing information [35]. Cohesion is quantified by two primary metrics, 'Density' and 'Average Degree' [9]. We use the metrics for evaluating the cohesion and the formula are as follow:

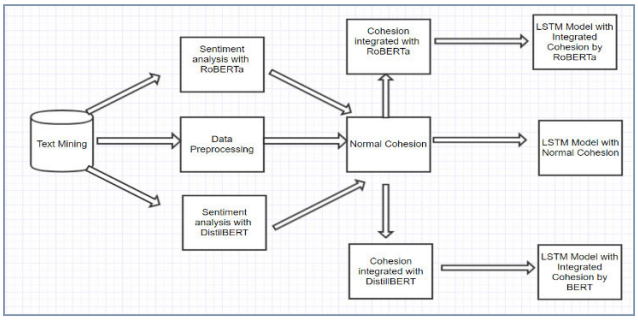

FIGURE 2. Research flowchart.

$$\text{Density} = \frac{\text{Actual Number of Connections}}{\text{NodeCount} \times (\text{NodeCount} - 1)/2}$$

FIGURE 3. Equation: density cohesion formula.

$$\text{Average Degree} = \frac{\text{Total Number of Quotations}}{\text{PostCount}}$$

FIGURE 4. Equation: average degree formula.

IV. RESULTS

Data was collected from the page discussing and predicting Bitcoin price movements in the economics-speculation section of Bitcointalk.com. The collected text data amounted to a total of 190,796 (Figure 5).

	Text	date
0	Quote from: Dafar on January 01, 2017, 05:11:5...	January 01, 2017, 05:11:51 PM
1	c'mon bitcoin, you can do it !!!\ntriple digit...	January 01, 2017, 05:11:51 PM
2	That 1k wall @Stamp is the only thing holding ...	January 01, 2017, 05:11:51 PM
3	Quote from: strawbs on January 01, 2017, 05:17...	January 01, 2017, 05:17:37 PM
4	And look at the volume. This is not rally leve...	January 01, 2017, 05:17:37 PM
...	...	...
190792	buddy sandwich	September 16, 2023, 11:50:35 PM
190793	Quote from: philipma1957 on September 16, 2023...	September 16, 2023, 11:50:35 PM
190794	Explanation\nChartbuddy thanks talkimg.com	September 16, 2023, 11:50:35 PM
190795	Closely clocking the growing difficulty.\n\nM...	September 16, 2023, 11:50:35 PM
190796	Explanation\nChartbuddy thanks talkimg.com	September 16, 2023, 11:50:35 PM

FIGURE 5. Result of text mining.

The collected data was converted into daily data by calculating the average of the corresponding date using the Groupby function. The converted daily data was analyzed by adding the variables of the existing cohesion (Figure 6) and the proposed model cohesion (Figure 7).

The graph (Figure 9) shows the residuals obtained from a Long Short-Term Memory (LSTM) model using the Average Cohesion Method. There are significant residuals at the beginning, which decrease progressively, indicating an improvement in model accuracy as more data is processed.

Figure 10 shows the residuals resulting from applying sentiment analysis with RoBERTa to cohesion measurements using the Average Cohesion Method. Initially, there are high residuals, which decrease over time, indicating that the integration of sentiment analysis with RoBERTa improves the accuracy of cohesion predictions as more data is processed.

	BCP	BTV	GCP	VIX	S&P	Cohe
0	995.4	41150.0	1151.73	18.50	2245.23	0.344262
1	1017.0	64950.0	1151.73	18.04	2243.25	0.297101
2	1033.3	54790.0	1162.00	12.85	2252.50	0.520548
3	1135.4	156270.0	1165.30	11.85	2264.25	0.400000
4	989.3	240010.0	1181.30	11.67	2264.25	0.409201
...	...	...	...	...	...	...
1829	26524.7	53820.0	1932.80	12.82	4555.00	0.428571
1830	26601.0	36470.0	1946.20	13.79	4498.00	0.305085
1831	26562.0	18450.0	1946.20	13.79	4498.00	0.440000
1832	26529.1	16570.0	1946.20	13.79	4498.00	0.424242
1833	26763.5	63350.0	1953.40	14.00	4501.50	0.406250

1834 rows × 6 columns

FIGURE 6. Result of cohesion (BCP: Bitcoin Closing price, BTV: Bitcoin Trading Volume, GCP: Gold Closing Price, VIX: Volatility Index, S&P: S&P Index, Cohe: Cohesion Quality).

	BCP	BTV	GCP	VIX	S&P	Cohe
0	995.4	41150.0	1151.73	18.50	2245.23	0.412133
1	1017.0	64950.0	1151.73	18.04	2243.25	0.132516
2	1033.3	54790.0	1162.00	12.85	2252.50	-0.283069
3	1135.4	156270.0	1165.30	11.85	2264.25	-0.288518
4	989.3	240010.0	1181.30	11.67	2264.25	-0.272094
...	...	...	...	...	...	...
1829	26524.7	53820.0	1932.80	12.82	4555.00	0.626292
1830	26601.0	36470.0	1946.20	13.79	4498.00	0.555471
1831	26562.0	18450.0	1946.20	13.79	4498.00	0.588218
1832	26529.1	16570.0	1946.20	13.79	4498.00	0.733063
1833	26763.5	63350.0	1953.40	14.00	4501.50	0.763426

1834 rows × 6 columns

FIGURE 7. Result of proposed LSTM cohesion (BCP: Bitcoin Closing price, BTV: Bitcoin Trading Volume, GCP: Gold Closing Price, VIX: Volatility Index, S&P: S&P Index, Cohe: Cohesion Quality).

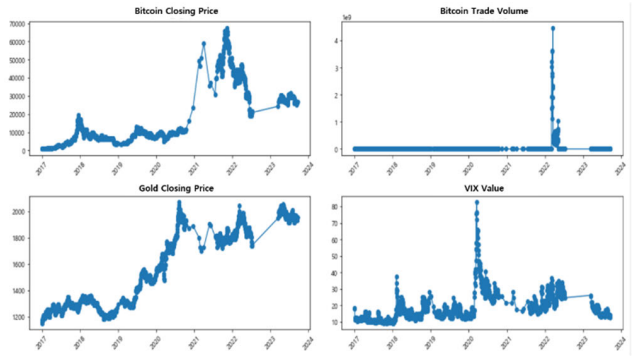

FIGURE 8. Result of each variables.

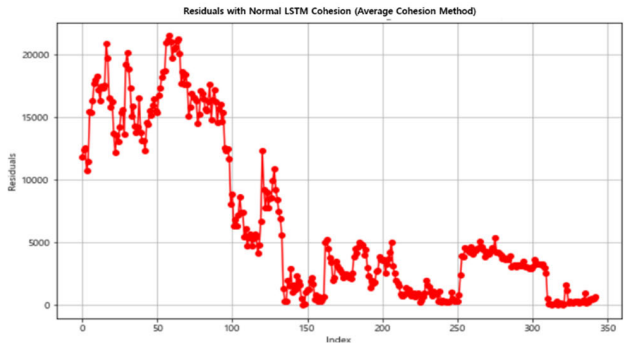

FIGURE 9. Result of residuals with LSTM.

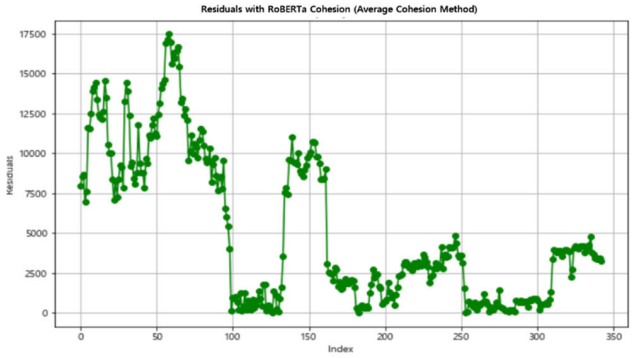

FIGURE 10. Result of residuals with RoBERTa.

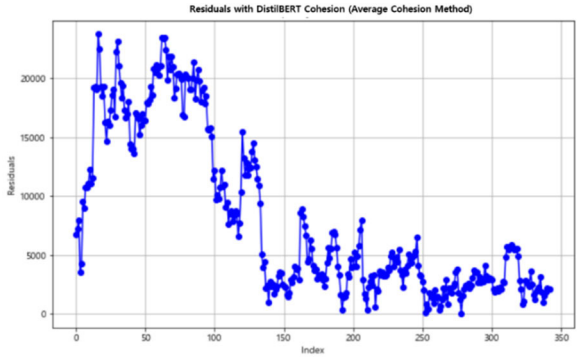

FIGURE 11. Result of residuals with distill-bert integrated model.

Initially, the residuals are quite high, peaking around 20,000, indicating less accurate predictions. As the index increases, the residuals show a noticeable decreasing trend, dropping significantly to below 5,000 by index 100, suggesting improved prediction accuracy. Between indices 100 and 300, the residuals fluctuate but remain lower than the initial values, indicating varied yet improved accuracy. After index 300, the residuals stabilize further, mostly below 5,000, showing that the model reaches a relatively steady state of accuracy. Overall, the graph demonstrates that integrating DistilBERT-based sentiment analysis with cohesion measurements into the LSTM model leads to an initial period of high residuals, followed by significant improvement and stabilization in prediction accuracy as the model processes more data.

The performance indicators of the existing model and the proposed model were compared and analyzed using Mean Square Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) indicators. The values derived from the LSTM analysis reflecting the existing cohesion were MSE: 47,685,990, RMSE: 6,905.50, MAE: 5,065.4, MAPE: 34.94%, and the MSE value derived from the LSTM applied with the RoBERTa model was 45,880,683, RMSE: 6773.52, MAE: 5364.72, MAPE:30.82%. In the case of the DistilBERT model, the values derived were MSE: 1,839,094, RMSE: 1,356.13, MAE: 1,078.54, MAPE: 16.24% (Table 1).

TABLE 1. Results of model performance.

	Performance Measure		
	LSTM	RoBERTa	DistilBERT
MSE	47,685,900	45,880,683	1,839,094
RMSE	6,905.50	6,773.52	1,356.13
MAE	5,065.40	5,364.72	1,078.54
MAPE	34.94%	30.82%	16.24%

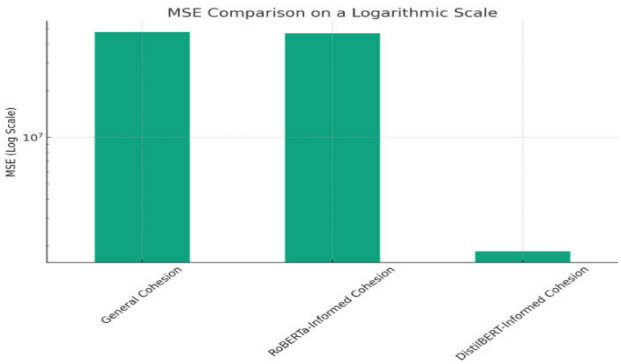

FIGURE 12. Result of MSE comparison.

The chart illustrates that while general cohesion and RoBERTa-informed cohesion models perform similarly, the DistilBERT-informed cohesion model significantly outperforms them, achieving much lower MSE and thus providing more accurate predictions. This highlights the superior predictive power of the DistilBERT-informed approach in the context of this analysis.

The chart compares the Root Mean Squared Error (RMSE) of three different models: General Cohesion, RoBERTa-Informed Cohesion, and DistilBERT-Informed Cohesion. The x-axis represents the models, while the y-axis shows the RMSE values. The General Cohesion model has an RMSE of around 7,000, similar to the RoBERTa-Informed Cohesion model, indicating comparable prediction.

The figure 14 compares the Mean Absolute Error (MAE) of three different models: General Cohesion, RoBERTa-Informed Cohesion, and DistilBERT-Informed Cohesion. The x-axis represents the models, while the y-axis shows the MAE values. The General Cohesion

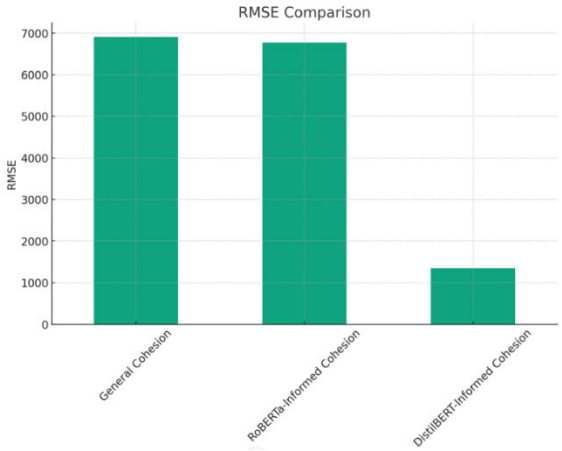

FIGURE 13. Result of RMSE comparison.

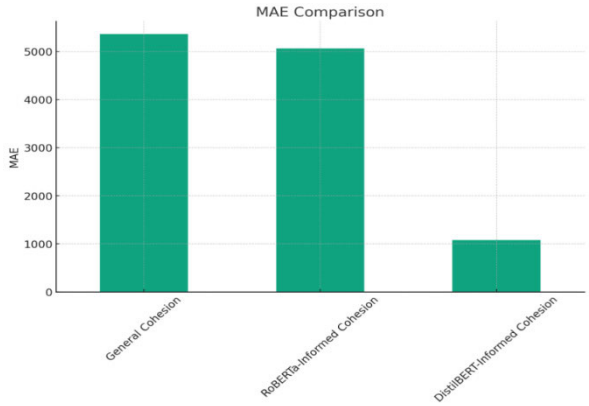

FIGURE 14. Result of MAE.

model has an MAE of approximately 5,000, similar to the RoBERTa-Informed Cohesion model, indicating comparable performance in terms of absolute prediction errors. In contrast, the DistilBERT-Informed Cohesion model significantly outperforms the other two, with an MAE of around 1,000, demonstrating much higher prediction accuracy and lower absolute errors. This comparison highlights the superior performance of the DistilBERT-Informed Cohesion model.

The graph (Figure 15) compares the Mean Absolute Percentage Error (MAPE) of three different models: General Cohesion, RoBERTa-Informed Cohesion, and DistilBERT-Informed Cohesion. The x-axis represents the models, while the y-axis shows the MAPE values. The General Cohesion model has the highest MAPE at around 35%, followed closely by the RoBERTa-Informed Cohesion model at approximately 31%, indicating that both models have relatively high percentage errors. The DistilBERT-Informed Cohesion model, with a MAPE of around 16%, significantly outperforms the other two, indicating more accurate predictions in terms of percentage error. This comparison underscores the enhanced predictive capability of the DistilBERT-Informed Cohesion model.

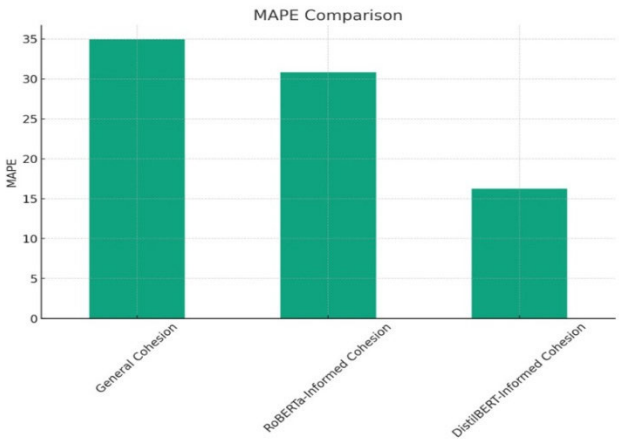

FIGURE 15. Result of MAPE comparison.

The analysis results showed that the DistilBERT model had the best performance, followed by RoBERTa and LSTM reflecting the existing cohesion variables. The residuals of each model were used to compare the performance between models. The paired sample t-test results showed that both RoBERTa and DistilBERT improved prediction performance compared to the existing LSTM. Model 1 is the result of comparing and analyzing the existing LSTM and RoBERTa models. The t-stat was 6.0620, which was significant at the 99% level. The analysis of the existing LSTM compared with DistilBERT (Model 2) also showed that the DistilBERT model predicts accurate result values (Table 2).

TABLE 2. Results of T-test.

	T_Stat	P_Value
Model 1 (LSTM and RoBERTa)	6.0620	.000
Model 2 (LSTM and DistilBERT)	19.5002	.000

V. CONCLUSION

This study conducts to enhance the prediction of the cryptocurrency market by extending the LSTM. The results of this study are as follows. First, it is confirmed that the psychological factors of consumers, which are mainly used in previous stock market prediction analysis studies, can also be used in the cryptocurrency market. The RoBERTa and DistilBERT models used in this study are confirmed to be appropriate models to increase the prediction accuracy of the cryptocurrency market (H1). If the consumer psychology applied to the LSTM is simplified or excluded, it means that the accuracy can decrease from the perspective of market prediction. In other words, consumer sentiment analysis is essential for accurate market prediction.

Second, as a result of comparing the LSTM and the RoBERTa model, it is judged that the prediction of the

RoBERTa model is better, which is based on the characteristics of the consumer's psychology being subdivided (−1, 0, 1). These results mean that the more the consumer's psychology is subdivided, the more the accuracy of market prediction can be improved. Therefore, it is necessary to apply a model that subdivides consumer sentiment for the future improvement of the accuracy of cryptocurrency market prediction.

Third, the DistilBERT model, which specializes in qualitative analysis of language by assigning sentiment labels not only to words but also to sentences, shows much better market prediction than the LSTM and RoBERTa model. For high accuracy of text analysis, it means that the analysis model needs to understand the flow of context including words rather than focusing on analyzing words. The fact that the DistilBERT model, which is divided into −1 and 1, shows higher accuracy than the RoBERTa model divided into −1, 0, 1, can be evidence to support these results. For effective text analysis, it is necessary to select an appropriate analysis model, and especially for sentiment analysis, it is judged that a model specialized in qualitative analysis including the flow of context is needed.

The implications of this study have several critical points. In the realm of big data analysis for market prediction, researchers need to consider not only the volume of data but also its qualitative dimensions. Beyond mere word frequency, grasping the contextual coherence of text holds pivotal importance. Neglecting linguistic subtleties in text analysis can potentially yield outcomes that deviate from actuality. Moreover, this study underscores the essentiality of integrating investors' psychological factors into the analysis of cryptocurrency markets. This study confirmed that within a single community, the coherence can significantly influence both investor psychology and investment decisions.

However, the research is limited by its failure to distinguish between psychological cohesion and information cohesion. Therefore, future studies are expected to achieve more precise outcomes by developing models or methodologies capable of quantifying each aspect of cohesion during analysis.

ACKNOWLEDGMENT

The authors do not have any conflicts of interest. *(Namjae Cho, Jae Hyun Byun, and Giseob Yu contributed equally to this work.)*

REFERENCES

[1] S. A. Sarkodie, M. Y. Ahmed, and P. A. Owusu, "COVID-19 pandemic improves market signals of cryptocurrencies–evidence from Bitcoin, Bitcoin cash, Ethereum, and Litecoin," *Finance Res. Lett.*, vol. 44, Jan. 2022, Art. no. 102049, doi: 10.1016/j.frl.2021.102049.

[2] H. Rezaei, H. Faaljou, and G. Mansourfar, "Stock price prediction using deep learning and frequency decomposition," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114332, doi: 10.1016/j.eswa.2020.114332.

[3] H. R. Kim, S. H. Hong, and H. Hong, "Machine learning based stock price fluctuation prediction models of KOSDAQ-listed companies using online news, macroeconomic indicators, financial market indicators, technical indicators, and social interest indicators," *J. Korea Multimed. Soc.*, vol. 24, no. 3, pp. 448–459, Mar. 2021, doi: 10.9717/KMMS.2020.24.3.448.

[4] S. Aras, "Stacking hybrid GARCH models for forecasting Bitcoin volatility," *Expert Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114747, doi: 10.1016/j.eswa.2021.114747.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[7] Z. Hu, Y. Zhao, and M. Khushi, "A survey of forex and stock price prediction using deep learning," 2021, *arXiv:2103.09750*.

[8] E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," *Expert Syst. Appl.*, vol. 129, pp. 273–285, Sep. 2019, doi: 10.1016/j.eswa.2019.03.029.

[9] P. Xie, H. Chen, and Y. J. Hu, "Signal or noise in social media discussions: The role of network cohesion in predicting the Bitcoin market," *J. Manage. Inf. Syst.*, vol. 37, no. 4, pp. 933–956, Dec. 2020, doi: 10.1080/07421222.2020.1831762.

[10] H. Purohit, Y. Ruan, D. Fuhry, S. Parthasarathy, and A. Sheth, "On the role of social identity and cohesion in characterizing online social communities," 2012, *arXiv:1212.0141*.

[11] A. Kumar and E. Operti, "Missed chances and unfulfilled hopes: Why do firms make errors in evaluating technological opportunities?" *Strategic Manage. J.*, vol. 44, no. 13, pp. 3067–3097, Dec. 2023, doi: 10.1002/smj.3543.

[12] S. González-Bailón and Y. Lelkes, "Do social media undermine social cohesion? A critical review," *Social Issues Policy Rev.*, vol. 17, no. 1, pp. 155–180, Jan. 2023, doi: 10.1111/sipr.12091.

[13] M. S. Mizruchi and L. B. Stearns, "Getting deals done: The use of social networks in bank decision-making," *Amer. Sociol. Rev.*, vol. 66, no. 5, p. 647, Oct. 2001, doi: 10.2307/3088952.

[14] T. L. Huston and G. Levinger, "Interpersonal attraction and relationships," *Annu. Rev. Psychol.*, vol. 29, no. 1, pp. 115–156, Jan. 1978, doi: 10.1146/annurev.ps.29.020178.000555.

[15] X. Guang, Y. Gao, P. Liu, and G. Li, "IMU data and GPS position information direct fusion based on LSTM," *Sensors*, vol. 21, no. 7, p. 2500, Apr. 2021, doi: 10.3390/s21072500.

[16] M. A. Nadif, Md. T. Rahman Samin, and T. Islam, "Stock market prediction using long short-term memory (LSTM)," in *Proc. 2nd Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Apr. 2022, pp. 1–6, doi: 10.1109/ICAECT54875.2022.9807655.

[17] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. C. Khatri, "Predicting stock market index using LSTM," *Mach. Learn. Appl.*, vol. 9, Sep. 2022, Art. no. 100320, doi: 10.1016/j.mlwa.2022.100320.

[18] J. Singh, R. Thulasiram, and A. Thavaneswaran, "LSTM based algorithmic trading model for Bitcoin," *IEEE Xplore*, vol. 1, no. 1, pp. 344–351, Jan. 2022, doi: 10.1109/SSCI51031.2022.10022021.

[19] H. K. Andi, "An accurate Bitcoin price prediction using logistic regression with LSTM machine learning model," *J. Soft Comput. Paradigm*, vol. 3, no. 3, pp. 205–217, Sep. 2021, doi: 10.36548/jscp.2021.3.006.

[20] J. Lei and Q. Lin, "Analysis of gold and Bitcoin price prediction based on LSTM model," *Acad. J. Comput. Inf. Sci.*, vol. 5, no. 6, pp. 95–100, 2022, doi: 10.25236/AJCIS.2022.050614.

[21] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[22] S. Tam, R. B. Said, and Ö. Ö. Tanriöver, "A ConvBiLSTM deep learning model-based approach for Twitter sentiment classification," *IEEE Access*, vol. 9, pp. 41283–41293, 2021, doi: 10.1109/ACCESS.2021.3064830.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., Jun. 2017, pp. 6000–6010.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[26] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1–20.
[27] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, Aug. 2019.
[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[29] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2006, pp. 535–541, doi: 10.1145/1150402.1150464.
[30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv.1503.02531*.
[31] J. D. Brown and P. Duguid, "Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation," *Org. Sci.*, vol. 2, no. 1, pp. 40–57, Mar. 1991.
[32] T. Rowley, D. Behrens, and D. Krackhardt, "Redundant governance structures: An analysis of structural and relational embeddedness in the steel and semiconductor industries," *Strategic Manage. J.*, vol. 21, no. 3, pp. 369–386, Mar. 2000.
[33] B. Fang, Z. Zheng, Q. Ye, and P. B. Goes, "Social influence and monetization of freemium social games," *J. Manage. Inf. Syst.*, vol. 36, no. 3, pp. 730–754, Jul. 2019, doi: 10.1080/07421222.2019.1628878.
[34] H. J. Jeong and M. H. Kim, "HGGC: A hybrid group recommendation model considering group cohesion," *Expert Syst. Appl.*, vol. 136, pp. 73–82, Dec. 2019, doi: 10.1016/j.eswa.2019.05.054.
[35] S. Baquero, F. Montes, I. Stankov, O. L. Sarmiento, P. Medina, S. C. Slesinski, F. Diez-Canseco, M. F. Kroker-Lobos, W. T. Caiaffa, A. Vives, M. Alazraqui, T. Barrientos-Gutiérrez, and A. V. D. Roux, "Assessing cohesion and diversity in the collaboration network of the SALURBAL project," *Sci. Rep.*, vol. 13, no. 1, p. 7590, May 2023, doi: 10.1038/s41598-023-33641-x.

NAMJAE CHO received the Ph.D. degree in MIS from Boston University, USA. He is currently a Professor of MIS with the School of Business, Hanyang University, Seoul, South Korea. He has published research articles in journals, including *Industrial Management and Data Systems*, *Computers in Industry*, *International Journal of Information Systems and Supply Chain Management*, and *Data and Knowledge Engineering*. He also published several books, including *Supply Network Coordination in the Dynamic and Intelligent Environment* (IGI Global) and *Innovations in Organizational Coordination Using Smart Mobile Technology* (Springer, 2013). He consulted government organizations and several multinational companies. His research interests include technology planning and innovation, analysis of IT impacts, strategic alignment and IT governance, knowledge management and industrial ICT policy, design thinking, and the management of family business.

JAE HYUN BYUN was born in April 1994. He received the bachelor's degree in economics (psychology) from the University of Massachusetts Boston and the master's degree in business informatics from Hanyang University. Throughout his academic and professional journey, he has embarked on several projects that reflect passion and expertise in the field. These include: Predicting anomalies in manufacturing data using isolation random forest, developing an automated news clipping program using selenium, creating a program that expresses the temperature of text based on sentiment, using a sentiment analysis dictionary, developing an automated data collection program utilizing selenium, analyzing the strengths and weaknesses of the top two brands in the chicken breast industry using natural language processing of consumer reviews (TF-IDF and frequency analysis), predicting calorie content with generalized additive models (GAM), comparing the performance of LSTM and DNN models in time-series data prediction for chemical processes, data visualization based on the integration of SQL and Python programs, and improving LSTM prediction performance by incorporating qualitative aspects of cohesion using RoBERTa and DistillBERT applied to bitcoin forum data and modeling an image classifier using CNN and vision transformer (VIT). His diverse project experience showcases his ability to apply complex machine learning techniques and data analysis tools across a variety of domains, such as driving insights and innovation in each endeavor. His research interests include deep learning, natural language processing, and predictive modeling for image classification.

GISEOB YU received the bachelor's degree from Kangwon National University, the M.B.A. (Family Business Track) degree from Y.E.S., and the Ph.D. degree in MIS from Hanyang University. He was a Research Professor with Kyungpook National University. He is currently an Adjunct Professor with Hanyang University. His research interests include trend prediction utilizing big data, user experience analysis, entrepreneurship, and family business management, particularly in succession planning and digital transformation within family enterprises.