Sentiment analysis of users in the financial market using the BERT algorithm in the New York Stock market

Nima Heidari[1], Saeed Mirzamohammadi [2*], Babak Amiri [2]

[1] Ph.D. candidate, Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
[2] Assistant Professor, Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

*** Corresponding author email address**: mirzamohammadi@iust.ac.ir

Received: 2025-10-01 **Reviewed:** 2026-01-24 **Revised:** 2026-02-01 **Accepted:** 2026-02-07 **Published:** 2026-07-01

Abstract

Nowadays, analyzing individuals' sentiments is supposedly a main indicator in financial management and economic forecasts. Human analysis of different opinions and news during exchanging different opinions between individuals can take several minutes, and investors in financial markets need to make decisions quickly. Such challenging scenarios require faster ways for investors to make decisions. The current research analyzes the sentiments of users in the financial market using the BERT algorithm in the New York stock market based on profit news. This research has set up a data set based on the BERT model to evaluate and predict changes in New York stock market prices for Tesla and has analyzed the results through Python software. This research introduced two multi-class and multi-label models based on individuals' opinions to classify data on users' sentiments. The results showed that the multi-label model without considering the time lag with an F1 score equal to 0.846 has better performance compared to the multi-class models.

Keywords: *user sentiments, financial market, BERT algorithm, New York stock market*

How to cite this article:

Heidari, N., Mirzamohammadi, S., & Amiri, B. (2026). Sentiment analysis of users in the financial market using the BERT algorithm in the New York Stock market. Management Strategies and Engineering Sciences, 8(4), 1-11.

1. Introduction

An important part of the decision-making process is to take advantage of the opinions of individuals in different fields. The Internet and the Web now (and others) make it possible for us to use the opinions and experiences of a wide variety of individuals who are neither familiar to us nor recognized professional critics in our everyday decisions. Individuals and organizations are increasingly using public opinion in these media for their decision-making process with a dramatic growth of social media (eg, surveys, forum discussions, blogs, and social networks) on the Web. However, finding and monitoring tweets in forums and examining their information contents is still a difficult task because of the large variety of sites. An adult reader will have difficulty identifying relevant sites and accurately summarizing their information and opinions. Nowadays, customers and business owners use these opinions to identify the strengths and weaknesses of their products, but it is no longer possible to review them case by case and draw conclusions because of an increase in the volume of tweets. In conclusion, a system is necessary to acquire knowledge automatically under sentiment analysis (also known as opinion mining) [1].

Nowadays, sentiment analysis is used as a main element in various fields such as financial management, marketing, and forecasting economic changes in different countries. Market sentiment is a criterion that displays the sentiments of buyers and sellers. It is sometimes used as a method to analyze market prices and trader behavior. Some price movements or news have positive or negative effects on traders' and investors' sentiments. The very behavior and emotions of investors and traders shape market trends and prices. Now, the role of sentiment analysis is to identify positive, negative, and neutral sentiments in the market and give traders a general view of the future of the market. Sentiment analysis is a method of market analysis that predicts market behavior along with technical analysis and fundamental analysis. Activity in the financial markets

depends on the sentiments of the traders. These sentiments include fear, greed, worry, excitement, etc. Sentiment analysis is a modern method for analyzing financial markets, which is based on the analysis of traders' emotions and sentiments. This analytical method is usable in all financial markets, including the stock market, digital currency, forex, etc.

Most people turn to the stock market performance of countries as the best indicator of how well the economy is doing. Stock markets cover all industries in all sectors of the economy. This means that they use the cycle of the economy and the hopes and fears of the population that create growth and wealth as a barometer. Stock markets increase investment. Raising capital allows companies to grow their businesses, expand their activities, and create employment in the economy. This investment is a main driver of economic trade, growth, and prosperity. Stock markets offer investors a way to invest money to potentially gain a share of a company's profits. Financial markets, inclusive of stock markets, are markets that have always attracted the attention of many people. In conclusion, the social networks of these markets have received much attention. Users of these social networks publish their opinions as posts or tweets based on the state of the stock or the entire market. Users' opinions can be a reflection of the performance, news, or price value of companies and it can be completely emotional and non-technical without any analysis and support. Therefore, these comments contain positive or negative sentiments. Sentiment analysis is an evaluation of people's attitude towards each share, which is usable for senior managers in various organizations, and companies for whom the satisfaction or dissatisfaction of shareholders is important, or for predicting the future trend of a share.

The unique feature of these networks is that the price information of each share is available for that share on the day of commenting. As stocks are growing or falling every day, the comments that shareholders or analysts publish daily are affected by these stock growths and declines. Nowadays, researchers try to use artificial intelligence techniques such as convolutional neural networks to evaluate opinions, analyze emotional data of customers or users in virtual spaces, and identify and predict changes in the market. Intelligent sentiment analysis systems analyze and classify people's opinions, descriptions, and attitudes toward the raised subjects into positive, negative, and neutral groups. Therefore, the current research has analyzed users' sentiments in the financial market through the artificial intelligence-based BERT algorithm. BERT algorithm is a network-based intelligent algorithm that processes and explores pre-trained natural languages. Putting it simply, the BERT algorithm has focused on recognizing small words and phrases to provide better search results. The subject of this research is the New York stock market. The purpose of sentiment analysis is to automatically extract individuals' sentiments from social networks and text documents to improve performance accuracy in analyzing users' sentiments.

The second part of the research has reviewed the literature on the subject. The third part has introduced the method and collected data. The fourth part discusses the analysis of the results and the evaluation of the performance and accuracy of the presented model, and finally, the last part presents general conclusions and suggestions for future research.

2. Literature review

Various studies in recent years have dealt with user sentiment analysis. Sosa et al. (2019) used the BERT algorithm to perform sentiment analysis of news articles and provide relevant information for decision-making in the stock market [2]. Stock news articles are manually labeled as positive, neutral, or negative to fine-tune this powerful sentiment analysis model for the stock market. The data of this research amounts to 582 documents from several financial news sources. They prepared a BERT model on this data set and gained a 72.5% F-score. Li et al. (2020) examine the performance of natural language processing in financial sentiment classification [3]. They collected public opinion about US stocks from the website Stocktwits. The messages on this website reflect the views of investors on stocks. These messages are classified into positive or negative sentiments through a BERT language model. The experimental results showed that the pre-trained BERT model is well-tuned on the labeled sentiment data set and can recognize investor sentiment with an accuracy of more than 87.3%. Li et al. (2021) propose a new sentiment analysis model for Chinese stock research based on BERT [4]. This model relies on a pre-trained model to improve classification accuracy. As the results showed, their method has higher accuracy and F1 than TextCNN, TextRNN, Att-BLSTM, and TextCRNN.

Allaparty and Mishra (2021) examined the relative effectiveness of four sentiment analysis techniques: (1) an unsupervised vocabulary-based model using SentiWordNet, (2) a traditional supervised machine learning model using logistic regression, (3) a supervised deep learning model

using long-short-term memory (LSTM), and (4) Advanced supervised deep learning model using BERT [5]. The results show that all four sentiment analysis algorithms have relative efficiency and the deep learning algorithm under advanced BERT supervision is superior to other methods. Li et al. (2021) deal with sentiment analysis of investors in the stock market based on the (BERT) method [4]. First, they extracted the sentiment value from the online information published by the stock investor through the BERT model. Second, they weighted these sentiment values according to the investor sentiment index calculation. Finally, they analyzed the relationship between investor willingness and stock returns through a two-stage cross-sectional regression validation model. Experiments showed that investor sentiment in online reviews has a significant impact on stock returns. The BRET model used in this article could achieve an accuracy of 97.35% for investor sentiment analysis, which is better than both LSTM and SVM methods.

Jafarian et al. (2021) try in their research to improve the aspect-based sentiment analysis (ABSA) method on the Persian dataset [6]. This research shows the potential of using a pre-trained BERT model and exploiting the use of sentence pair input in an ABSA task. The results show that the use of the pre-trained Farsi-BERT model together with the natural language inference (NLI-M) for auxiliary sentences can increase the accuracy of the ABSA task by 91%. Liu et al. (2022) proposed an innovative ML method to predict the Chinese stock market trend by addressing investors' sentiments and attitudes, market rules, and the length of past market data [7]. First, investors' sentiments and attitudes were collected to measure them. Then, a non-stationary Markov chain (NMC) model was used to capture the dynamic transmission of market rules. The (BERT) method was chosen to predict the market situation at time t, and the long-short-term memory model (LSTM) too. The proposed algorithm achieves better results with a prediction accuracy of 61.77%, annual return of 29.25%, and maximum loss of -8.29% compared to the other prediction methods. The proposed model obtained also the lowest prediction error: mean square error (0.095), root mean square error (0.0739), mean absolute error (0.104), and mean absolute percent error (15.1%). In conclusion, the proposed market forecasting model can help investors to get more accurate market forecasting information. Lin et al. (2022) proposed a BERT-based model that uses ensemble learning methods with textual sentiment analysis to identify negative news [8]. The working model of the proposed system has two phases. The first phase collects negative news and creates a development model to analyze the correlation between text sentiment and negative news. The second phase identifies negative news by analyzing textual sentiments with the group learning technique and BERT model. The experimental results showed the F1 score of the proposed model reaches 66.3%, an increase of 7.8% compared to the Lagrange Support Vector Machine (LSVM) model. Ayapa et al. (2023) investigate the application of BERT and LSTM for stock price prediction through historical stock data and sentiment classification of Twitter posts [9]. The proposed model combines the power of LSTM to capture sequential patterns in stock prices and the text-understanding ability of BERT in tweet sentiment analysis. The experimental evaluation results showed that the proposed model surpassed conventional time series models and other deep learning models in accuracy and robustness. Jalodri et al. (2023) seek to comprehensively investigate deep BERT and CNN models used in sentiment analysis of COVID-19 tweets [10]. This study provides potential future research directions for the development of a high-quality BERT model. The results of this research have provided important insights into public attitudes, supporting decision-makers in understanding the overall perspective, identifying misinformation, and guiding emergency response tactics.

Nabila et al. (2023) carried out a test to identify comments containing negative sentences in social media in Indonesia through a previously trained model for Indonesians [11]. This study performed a multi-label classification and evaluated the classification results of Multilingual BERT (MBERT), IndoBERT, and Indo-Roberta Small models. The optimal result of this study is the use of the IndoBERT model with an F1 score of 0.8897. Hao et al. (2023) examined the method of recognizing and analyzing the text of the official sentiment document based on the neural network BERT model [12]. It extracts the sentiment information contained in the Internet textual information through the BERT-SVM model algorithm under deep learning and then extracts the user's sentiment. The sentiment analysis on the sentences of the article occurs by considering the influencing factors of individuals, society and even the country and puts the method in a position of analyzing the different sentiments a sentence or each word shows. The experiments, through training sentiments by LSTM-RNN and LSTM-RNN-word2vec models, show that the average accuracy of sentiment classification is 95.12%, the result of K-nearest neighbors is 90.87%, and the Bayesian classifier is 86.84%. Finally, as the comparison reveals, the BERT-SVM model improves the accuracy of text sentiment classification. Jain

et al. (2023) present a two-way encoder model based on bidirectional Encoder Representations from Transformers (BERT) that uses BERT as a pre-trained language model to generate word formation [13]. Three layers of Dilated Convolutional Neural Network (DCNN) overlaid with a global average pooling layer help fine-tune the model. Our implemented BERT-DCNN model performs dimensionality reduction and absorbs the associated dimensionality gain against any information loss. Furthermore, the model can capture long-term dependencies through different expansion rates. An emotional knowledge base is included in our model and enables it to achieve concept-level emotion analysis. Our empirical study shows the significance of the implemented model in F measurement, recall, accuracy, and precision with different machine learning models.

As the conducted studies show, the studies in recent years have focused on the problem of user sentiment analysis, but no research examined so far the user sentiment analysis in the stock market based on social networks of the stock market. Recent studies, despite the importance of the New York Stock Exchange, as an important component in the global economy, have not dealt with the impact of users' sentiments on this trading market. Therefore, this research examines the users' sentiment analysis in the financial market through an artificial intelligence-based BERT algorithm. The proposed model of this research, as an innovative issue, has been compared with the common models in other studies and has evaluated its performance and accuracy. Therefore, the innovations of the current research are:

i. Considering users' opinions in social networks of the stock market to analyze the sentiments of users based on profit news

ii. Analyzing user sentiments in the financial market through the artificial intelligence-based BERT algorithm

iii. Investigating the impact of users' opinions on stock prices and classifying them in multi-class classification and multi-label classification

3. Data and Software

3.1. Data

The present study has prepared a model based on the BERT artificial intelligence approach for evaluating and predicting changes in New York stock market prices for Tesla. Therefore, it will examine the price data of Tesla shares from the New York stock market in the period from 01/01/2020 to 11/13/2020, and the users' opinions on these shares from reliable sources to determine the analysis of the final price of this stock in the future day. Thus, it has labeled the analysis of users' opinions in four different ways. Historical data of Tesla stock price from January 16, 2020, to November 13, 2020 (318 days in total) including opening, closing, high and low prices were taken from www.yahoofinance.com.

3.2. Software

Identifying suitable features and analyzing users' opinions is done by the BERT algorithm in Python software. The accuracy of the results of the model has been specified with F1 and Accuracy indexes. Python is a high-level, "object-oriented" programming language with integrated dynamic semantics for web and application development. This programming language is highly attractive in the field of rapid development of application software.

4. Methodology

Our approach in this article is to use an unsupervised natural language processing model called the BERT algorithm. This algorithm can accurately implement 11 common cases of natural language processing after proper tuning and become a booster for natural language processing and understanding. This algorithm is deeply and precisely bidirectional (left and right of an expression or sentence). This means that the words before and after each word are properly checked and matched with the information contained in Wikipedia so that the algorithm can gain a correct understanding of an idea of what the user is looking for.

4.1. Preliminary review of expert opinions

The collected information are creating a database of the problem under study including the columns of the date and day of publication of the users' opinion and its full description, the opening, high, low, and closing prices, and 4 different labels to classify the articles. The following describes these four methods:

1) This label, which is calculated based on the percentage of the changes in the closing price compared to the opening on the day of the publication of the user's opinion, puts the data of the user's opinion into 3 groups.

2) This label, which is calculated based on the percentage of changes in the closing price on the day of publication of users' opinions and the closing price on the day before the

publication of users' opinions, puts the data of users' opinions into 3 groups.

3) This label, which is calculated based on the delta of the closing price on the day of publication of the users' opinion and the closing price on the day before the publication of the users' opinion, divided by the average indicator of the actual range on the day before the publication of the users' opinion, divides the data of the users' opinion into 3 groups.

4) This label is created by sentiment analysis and using the FinBert sentiment analysis model. All user comments are analyzed every day based on this model and labeled in 3 groups. The average weight of the labels of all users' opinions has been calculated and taken into account to determine the final tag of each day. Thus, we assigned a weight of zero for the opinion of users with the Bad label, a weight of 0.5 for the opinion of the users with the Neutral label, and a weight of 1 for the opinion of the users with the Good label, and then we calculated the average weight of the labels of all the user opinions published on that day. If the average weight of users' opinion is in the range of 0 to 0.33, the overall label of that day is bad, if it is in the range of 0.33 to 0.66, the overall label of that day is neutral, and if it is in the range of 0.66 to 1, the overall label will be good.

Table 1 shows the grouping and the frequency of the user comments data of each group based on the corresponding label.

Table 1. Grouping and the number of user opinion data of each group based on the first label

Label	Group	Group selection	Observations
Label 1	Bad	$\left(\frac{close}{open}\right) - 1 \leq -3٪$	51
	Neutral	$-3٪ < \left(\frac{close}{open}\right) - ١ < 3٪$	178
	Good	$3٪ \leq \left(\frac{close}{open}\right) - 1$	52
Label 2	Bad	$\left(\frac{close_{today}}{close_{yesterday}}\right) - 1 \leq -3٪$	42
	Neutral	$-3٪ \leq \left(\frac{close_{today}}{close_{yesterday}}\right) - 1 < 3٪$	187
	Good	$3٪ \leq \left(\frac{close_{today}}{close_{yesterday}}\right) - 1$	52
Label 3	Bad	$\left(\frac{close_{today} - close_{yesterday}}{ATR_{yesterday}}\right) < -1٪$	23
	Neutral	$-1 \leq \left(\frac{close_{today} - close_{yesterday}}{ATR_{yesterday}}\right) < 1$	218
	Good	$1 \leq \left(\frac{close_{today} - close_{yesterday}}{ATR_{yesterday}}\right)$	40
Label 4	Bad	Average weight of news is in the range of 0 to 0.33	58
	Neutral	Average weight of news is in the range of 0.33 to 0.66	158
	Good	Average weight of news is in the range of 0.66 to 1	65

4.2. FinBERT model

The use of deep learning and pre-trained models in financial text mining is often unsuccessful because of the lack of labeled training data in this field. Therefore, the FinBERT model, which was a developed BERT method, was introduced in 2020 by Zhuang et al. [14] and was trained and published on a large amount of financial data published by large and specialized financial companies. This work has also used this model because of its better results compared to the original BERT model in financial text mining issues.

This article has used the FinBERT model to investigate the effect of users' opinions on stock prices and to classify users' opinions in different classes in the following two different ways:

1. Multi-class: The model in this mode is trained on the first to third labels (labels of the share price) separately, and the results and accuracy of each one are presented separately.

2. Multi-label: All the available labels for each user's comment in this mode are simultaneously placed in a binary vector. The main problem in this method is to find a model that maps inputs x to binary vectors y (assigning a value of 0 or 1 to each element (label) in y). For example, if the values of the first to fourth labels for a specific user comment are equal to "Good", "Neutral", "Good" and "Bad", the resulting vector in the multi-label model is:

(1,0,0,0,1,0,1,0,0,0,0,1)

4.3. Model evaluation criteria

This research has used the criteria of accuracy and F1 score to evaluate the presented models and compare their performance under different conditions. The accuracy criterion is chosen because of ease of understanding and use. Of course, it is noteworthy that this criterion is not usually used for multi-label models. The reason is that the accuracy calculation is a matter of relative taste in multi-label models because of the several different labels and existing displacements. It is important to calculate the accuracy for each label separately, to consider the pure accuracy of all labels, or even to evaluate a proportion of the accuracy of the labels. Therefore, this criterion is used less in multi-label models.

5. Results

All models in this section were trained once in normal mode, i.e. without time lag, and once again with a one-day time lag, and the results were analyzed to investigate the impact of news release time on share price changes.

5.1. Training models

We must first separate the input and output data to the model to train the described models. The input data to the model is a list of each of the elements of the Headline column beside each of the elements of the Article column. The output of the model is a list of 3 classes good, bad, and neutral, each of which has a subscript. Thus, if the value of the label is good, its subscript will be zero, if the value of the label is bad, its subscript will be one, and if the value of the label is neutral, its subscript will be 2. Then these data are separated based on the test and training size equal to 20% to 80% and in shuffle. We tokenize the test and training data, and then the model training process starts by calling it.

5.2. Multi-3class model outputs

This section presents separately the results of the training of the introduced model on different labels. All FinBERT models in this section have been trained up to 10 epochs to evaluate the model and the trend of its accuracy changes. It is noteworthy that we refrained in this section from presenting the results of model training on the fourth label, which was prepared based on the analysis of users' sentiments, because of the lack of added value in the predictability of this label. Indeed, even if we assume that the results of this model have good accuracy and performance, it will not help us in predicting the share price the next day because this label has nothing to do with the share price and only expresses the news-motivated sentiments of that day.

5.3. Results for the first label without considering the time lag

Table 2 gives the output of this model, which is calculated by the percentage of changes in the closing price compared to the opening on the day of the news release, without considering the one-day time lag for the FinBERT model. As it is clear, the amount of loss resulting from the FinBERT model in the validation part up to the third epoch is decreasing and is increasing from the third epoch onwards, and this means an overfitting of the model and a false increase of the accuracy from the third epoch onwards. In conclusion, the third epoch is the basis for evaluating this model. As you can see, the accuracy of this model is 79.06% and the F1 score is 0.688.

Table 2. Results of FinBERT model training on the first label without considering the time lag

Epoch	Validation F1	Train F1	Validation Accuracy	Train Accuracy	Validation Loss	Train Loss
1	0.666613	0.642538	78.80829	75.123217	0.338377	0.396679
2	0.698258	0.672803	79.326425	78.171206	0.338036	0.360494
3	0.688967	0.697719	79.067358	79.079118	0.335994	0.350466
4	0.685371	0.727215	78.80829	80.505837	0.350019	0.340997
5	0.690029	0.73773	78.80829	80.376135	0.361398	0.335302
6	0.698856	0.761405	79.067358	81.54345	0.371598	0.330712
7	0.713495	0.772527	80.103627	82.191958	0.388799	0.324801
8	0.722321	0.781688	80.621762	82.905318	0.406881	0.318089
9	0.725447	0.786531	80.362694	83.294423	0.416292	0.311521
10	0.717956	0.807435	79.585492	84.785992	0.426971	0.303818

5.4. Results for the second label without considering the time lag

Table 3 presents the output of this model, which is calculated based on the percentage of changes in the closing price on the day of the news release with the closing price on the day before the news release, without considering the one-day time lag for the FinBERT model. As it is clear, the amount of loss resulting from this model in the validation part increases from the second epoch; so the first epoch is the basis for the evaluation of this model. As you can see, the accuracy of the model is 83.21% and the F1 score is 0.775.

Table 3. Results of FinBERT model training on the second label without considering the time lag

Epoch	Validation F1	Train F1	Validation Accuracy	Train Accuracy	Validation Loss	Train Loss
1	0.7754914	0.6609231	83.2124352	76.2905318	0.35015372	0.3838332
2	0.74549173	0.6948343	80.10362694	79.2088197	0.36261114	0.3461848
3	0.74161548	0.7250324	78.5492228	80.6355383	0.3584517	0.3350237
4	0.74357803	0.7466308	79.32642487	81.6083009	0.35111427	0.325636
5	0.7461751	0.7629181	79.84455959	82.3216602	0.35366659	0.3187976
6	0.74116395	0.7727859	78.80829016	82.6459144	0.35884066	0.3151649
7	0.74977459	0.7862847	80.3626943	83.4241245	0.36188346	0.3098916
8	0.73796648	0.7875875	78.03108808	83.0998703	0.37348839	0.3047847
9	0.72675628	0.7930631	75.95854922	83.2944228	0.38960819	0.2993594
10	0.7226851	0.8084087	75.44041451	84.3968872	0.40197894	0.2922463

5.5. Results for the third label without considering the time lag

Table 4 provides the output of this model, which is calculated based on the delta of the closing price on the day of the news release and the closing price on the day before the news release, divided by the average indicator of the actual range on the day before the news release, without considering the one-day time lag. As it is clear, the amount of loss of this model also increases in the validation part from the second epoch; so the first epoch is the basis for evaluating this model. As you can see, the accuracy of the model is 91.72% and the F1 score is 0.822. It is noteworthy that we have not provided, from this section onwards, the results of the BERT model training because of the weakness of the results.

Table 4. Results of model training on the third label without considering the time lag

Epoch	Validation F1	Train F1	Validation Accuracy	Train Accuracy	Validation Loss	Train Loss
1	0.822831	0.7676493	91.7253886	86.92607	0.2718267	0.297258
2	0.8223436	0.8109056	88.39378238	89.714656	0.2766144	0.2617956
3	0.8277789	0.8262115	88.65284974	90.103761	0.2804755	0.2520991
4	0.8157715	0.8396939	86.83937824	90.817121	0.2853489	0.2434021
5	0.819269	0.8456256	87.0984456	90.817121	0.293588	0.2366425
6	0.8232611	0.8561951	86.83937824	91.595331	0.3099501	0.231297
7	0.8227298	0.8710541	85.54404145	92.30869	0.3251992	0.22545
8	0.8118056	0.8834828	83.98963731	92.892348	0.3348161	0.2172837
9	0.8006229	0.8928977	81.91709845	93.670558	0.3459997	0.2091004
10	0.8037375	0.8989017	81.65803109	93.929961	0.3585061	0.2010751

5.6. Results for the first label considering one day of time lag

Table 5 presents the output of this model considering the time lag of one day. As it is clear, the amount of loss resulting from this model in the validation part increases from the fourth epoch, so the third epoch is the basis for the evaluation of this model. As you can see, the accuracy of the model is 75.96% and the F1 score is 0.6814.

Table 5. Results of model training on the first label considering a one-day time lag

Epoch	Validation F1	Train F1	Validation Accuracy	Train Accuracy	Validation Loss	Train Loss
1	0.6703282	0.6907875	76.735751	80.830091	0.3585845	0.3676341
2	0.6521973	0.7011396	75.958549	81.608301	0.3568041	0.3364252
3	0.681404	0.7273983	75.96321	82.840467	0.3536559	0.3269819
4	0.6825197	0.7391335	74.404145	82.581064	0.371321	0.318584
5	0.6735074	0.7530258	73.88601	82.775616	0.388744	0.3152119
6	0.6769119	0.7621104	74.663212	82.645914	0.4079304	0.3113212
7	0.6760955	0.7829373	74.145078	83.81323	0.4145566	0.3081656
8	0.6876913	0.7883828	74.663212	83.942931	0.4276242	0.304963
9	0.6813625	0.7990494	73.367876	84.461738	0.4271822	0.2992634
10	0.6754202	0.7085788	71.813472	85.175097	0.4272819	0.2925999

5.7. Results for the second label considering one day of time lag

Table 6 gives the output of this model considering the time lag of one day. As it is clear, the amount of loss resulting from this model in the validation part increases from the third epoch; so the second epoch is the basis for the evaluation of this model. As you can see, the accuracy of the model is 78.29% and the F1 score is 0.635.

Table 6. Results of model training on the second label considering a one-day time lag

Epoch	Validation F1	Train F1	Validation Accuracy	Train Accuracy	Validation Loss	Train Loss
1	0.6258021	0.6859886	77.772021	78.560311	0.3694396	0.3784268
2	0.6352464	0.7070768	78.290155	81.738003	0.3687085	0.3447171
3	0.654899	0.718904	78.03109	81.997406	0.3714403	0.3338749
4	0.6491474	0.7301781	77.253886	82.256809	0.3745548	0.325996
5	0.6634237	0.7444451	77.772021	82.710765	0.3783104	0.3199449
6	0.6742015	0.7660735	78.549223	83.424125	0.3829518	0.3121434
7	0.6638768	0.7804115	76.994819	84.007782	0.3908426	0.3017951
8	0.6605105	0.8026458	76.994819	85.499351	0.4037426	0.2923263
9	0.6562378	0.8069982	76.217617	85.564202	0.4200127	0.2850283
10	0.6554298	0.8137111	75.699482	85.953307	0.4356667	0.2787136

5.8. Results for the third label considering one day of time lag

Table 7 presents the output of this model considering the time lag of one day. As it is clear, the amount of loss resulting from this model in the validation part increases from the second epoch; so the same initial epoch is the basis for the evaluation of this model. As you can see, the accuracy of the model is 91.53% and the F1 score is 0.830.

Table 7. Results of model training on the third label considering a one-day time lag

Epoch	Validation F1	Train F1	Validation Accuracy	Train Accuracy	Validation Loss	Train Loss
1	0.83063	0.8483793	91.53886	93.281453	0.2675466	0.2487409
2	0.8444658	0.8636007	91.502591	95.032425	0.2823489	0.2165519
3	0.8289359	0.8680023	89.170984	94.773022	0.2916725	0.207459
4	0.8334721	0.8797535	89.170984	95.162127	0.2954153	0.1968077
5	0.8444621	0.8923452	89.689119	95.551232	0.2965869	0.18836
6	0.8380232	0.896081	88.134715	95.291829	0.3073274	0.1832566
7	0.836588	0.9033059	88.134715	95.810636	0.3229505	0.1760055
8	0.8391093	0.9107937	89.170984	96.199741	0.348344	0.1688942
9	0.8245823	0.9224503	87.357513	96.977951	0.3766214	0.1613287
10	0.8183235	0.9337444	86.321244	97.885863	0.3847452	0.1552799

5.9. Results of the outputs of the multi-label model

The results of the multi-label model in this section are presented once without considering the time lag and once again with the consideration of the one-day time lag. This model considers the effect of all labels simultaneously. All the models in this section have been trained up to 20 epochs to evaluate the model and the trend of its accuracy changes. It is noteworthy that the accuracy evaluation criterion, as explained in the previous chapter, does not have a good performance for multi-label models, so we have limited ourselves in this section to present the results of the F1 score criterion.

5.10. Results for the multi-label model without considering the time lag

Table 8 presents the output of this model. As it is clear, the amount of loss resulting from this model in the validation part increases from the second epoch; so the same initial epoch is the basis for the evaluation of this model. As you can see, the F1 score for this model is equal to 0.846.

Table 8. Results of training the multi-label model without considering the time lag

Epoch	Validation F1	Train F1	Validation Loss	Train Loss
1	0.846252158	0.844353624	0.409063196	0.408131159
2	0.847348489	0.871530469	0.41039062	0.376546806
3	0.846236378	0.881213688	0.413517418	0.369528339
4	0.85688391	0.889346123	0.423024662	0.364134793
5	0.850999754	0.896992855	0.43799111	0.359348158
6	0.858442879	0.908967587	0.449839145	0.355012413
7	0.856483922	0.925996833	0.456306552	0.350747954
8	0.859064956	0.927265827	0.462359748	0.347014601
9	0.858801275	0.935542001	0.46522078	0.343471355
10	0.860017629	0.941365111	0.468949928	0.338750809
11	0.856597842	0.946639504	0.472407165	0.333890834
12	0.85645512	0.949241117	0.473942442	0.329141486
13	0.861331717	0.957165187	0.47468871	0.324418659
14	0.877704603	0.955000743	0.47575175	0.319736015
15	0.898999027	0.954601073	0.480051829	0.314941681
16	0.897746674	0.958085336	0.486596401	0.310250621
17	0.897402775	0.962289527	0.4896629	0.30637083
18	0.888861916	0.96470784	0.491891158	0.303485547
19	0.891353207	0.970479796	0.49807474	0.298483017
20	0.896898095	0.977102822	0.503586464	0.292499553

5.11. Results for the multi-label model considering the time lag

Table 9 presents the output of this model. As it is clear, the amount of loss resulting from this model in the validation part increases from the second epoch; so the same initial epoch is the basis for the evaluation of this model. As you can see, the F1 score for this model is equal to 0.787.

Table 9. Results of multi-label model training considering the time lag

Epoch	Validation F1	Train F1	Validation Loss	Train Loss
1	0.787372742	0.77293761	0.465180589	0.476222604
2	0.790772253	0.780292891	0.467086565	0.448352097
3	0.789092658	0.800309658	0.472902167	0.441936767
4	0.783298937	0.808247148	0.481499741	0.437280626
5	0.779778449	0.826812551	0.495247636	0.433197944
6	0.784495394	0.830874087	0.510746306	0.429993182
7	0.781837897	0.835283619	0.526218076	0.425941943
8	0.778597654	0.835499925	0.54231405	0.422470952
9	0.791640425	0.839779951	0.558362939	0.41924737
10	0.804174342	0.843163374	0.572874572	0.416174996
11	0.815498749	0.84593094	0.583431389	0.412981656
12	0.813488477	0.853087738	0.591076787	0.409293984

13	0.813114582	0.856158892	0.595724967	0.405211609
14	0.808634165	0.858519904	0.598614858	0.400918837
15	0.808239631	0.86256665	0.601682543	0.396453762
16	0.810380864	0.861784752	0.603695679	0.391658188
17	0.808452556	0.870026532	0.599181179	0.385860409
18	0.816064374	0.878406744	0.59009136	0.379958685
19	0.809309801	0.887100134	0.57722574	0.375096385
20	0.824666599	0.879795275	0.569141594	0.37080359

5.12. Comparison Study

Now, we can analyze the results as follows. As Table 10 on the results of the multiclass models shows, the multiclass model has the highest accuracy and F1 score on the third label without considering the time lag. As expected, the third label has a better performance than the first and second labels. The reason can be attributed to technical analysts' use of the average real range indicator to predict share price movements and target their profit limits. Indeed, as mentioned earlier, the average real range indicator predicts the potential of future price movements based on past and current market fluctuations, which is more accurate than price movements alone.

Table 10. Examining the results of multi-class models

Model	Accuracy	F1
Multi-class model on the first label without considering the time lag	79.06736	0.688967
Multi-class model on the second label without considering the time lag	83.21244	0.775491
Multi-class model on the third label without considering the time lag	91.72539	0.822831
Multi-class model on the first label considering the time lag	75.96321	0.681404
Multi-class model on the second label considering the time lag	78.29016	0.635246
Multi-class model on the third label considering the time lag	91.53886	0.82063

Table 10 also clarifies that the accuracy criteria and F1 score for all multiclass models without considering the time lag are higher than the mode of these models when one day of time lag is considered for them. The reason can also be attributed to the efficient market hypothesis. An efficient market is a market in which the published information affects the stock price at a high speed and the prices adjust themselves according to this information. This hypothesis states that, when special news about a share is published in news agencies, it is expected that the share price will react quickly to the published news and move up or down depending on whether the news is positive or negative. Therefore, the performance of all models has been better without considering a one-day interruption and time lag.

As Table 11 on the results of multi-label models shows, the multi-label model, like the multi-class models, has a higher accuracy without considering the time lag, which is relatively significant.

Table 11. Examining the results of multi-label models

Model	Accuracy	F1
Multi-label model without considering the time lag	-	0.846252158
Multi-label model with a time lag	-	0.787372742

Comparing the results of these models (see Tables 10 and 11) reveals that the model proposed in this research, i.e., the multi-label model, in its best result, which is equal to F1=0.846, is better than the best result of the multiclass model, which is equal to F1=0.822.

6. Conclusion and Future Research

The present study has prepared a model based on the BERT artificial intelligence approach for evaluating and predicting changes in New York stock market prices for Tesla. Thus, it trained the BERT algorithm, which is a pre-trained model, and FinBERT's algorithm, which is the BERT algorithm trained on financial data, on the news data of the shares of this company. Thus, all the news of Tesla shares in the period from January 16, 2020, to November 13, 2020 (318 days in total), including opening, closing, high and low prices, were collected from reliable news websites and labeled through four different labels. Finally, the introduced models were trained to classify the users' sentiments based

on the labels. This was done in two modes multi-class and multi-label models. The results showed that the multi-label model without considering the time lag performs better than the multi-label models with the time lag. The multi-label model performs better than the multi-class models without considering the time lag, and it can be used in other research and practical works with an F1 score equal to 0.846. Finally, some suggestions for future research: changing labels, using high or low prices instead of final prices, using other news sources such as tweets instead of economic news, changing the number of classes, and finally examining the stocks of other companies and other financial markets

Authors' Contributions

Nima Heidari: Conceptualization, Methodology, Software and Writing. Saeed Mirzamohammadi and Babak Amiri: Supervision, review and editing.

Acknowledgments

Authors thank all participants who participate in this study.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

References

[1] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-based vs. BERT-based sentiment analysis: A comparative study in Italian," *Electronics,* vol. 11, no. 3, p. 374, 2022, doi: 10.3390/electronics11030374.

[2] M. G. Sousa, K. Sakiyama, L. de Souza Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, "BERT for stock market sentiment analysis," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019, pp. 1597-1601, doi: 10.1109/ICTAI.2019.00231.

[3] M. Li, L. Chen, J. Zhao, and Q. Li, "Sentiment analysis of Chinese stock reviews based on BERT model," *Applied Intelligence,* vol. 51, pp. 5016-5024, 2021, doi: 10.1007/s10489-020-02101-8.

[4] M. Li, W. Li, F. Wang, X. Jia, and G. Rui, "Applying BERT to analyze investor sentiment in stock market," *Neural Computing and Applications,* vol. 33, pp. 4663-4676, 2021, doi: 10.1007/s00521-020-05411-7.

[5] S. Alaparthi and M. Mishra, "BERT: A sentiment analysis odyssey," *Journal of Marketing Analytics,* vol. 9, no. 2, pp. 118-126, 2021, doi: 10.1057/s41270-021-00109-8.

[6] H. Jafarian, A. H. Taghavi, A. Javaheri, and R. Rawassizadeh, "Exploiting BERT to improve aspect-based sentiment analysis performance on Persian language," in *2021 7th International Conference on Web Research (ICWR)*, 2021, pp. 5-8, doi: 10.1109/ICWR51868.2021.9443131.

[7] C. Liu, J. Yan, F. Guo, and M. Guo, "Forecasting the market with machine learning algorithms: an application of NMC-BERT-LSTM-DQN-X algorithm in quantitative trading," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 16, no. 4, pp. 1-22, 2022, doi: 10.1145/3564531.

[8] S. Y. Lin, Y. C. Kung, and F. Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," *Information Processing & Management,* vol. 59, no. 2, p. 102872, 2022, doi: 10.1016/j.ipm.2022.102872.

[9] Y. Ayyappa, B. V. Kumar, S. P. Priya, S. Akhila, T. P. V. Reddy, and S. M. Goush, "Forecasting Equity Prices using LSTM and BERT with Sentiment Analysis," in *2023 International Conference on Inventive Computation Technologies (ICICT)*, 2023, pp. 643-648, doi: 10.1109/ICICT57646.2023.10134443.

[10] J. H. Joloudari *et al.*, "BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets," *Social Network Analysis and Mining,* vol. 13, no. 1, p. 99, 2023, doi: 10.1007/s13278-023-01102-y.

[11] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "BERT base model for toxic comment analysis on Indonesian social media," *Procedia Computer Science,* vol. 216, pp. 714-721, 2023, doi: 10.1016/j.procs.2022.12.188.

[12] S. Hao, P. Zhang, S. Liu, and Y. Wang, "Sentiment recognition and analysis method of official document text based on BERT-SVM model," *Neural Computing and Applications,* pp. 1-12, 2023, doi: 10.1007/s00521-023-08226-4.

[13] P. K. Jain, W. Quamer, V. Saravanan, and R. Pamula, "Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis," *Journal of Ambient Intelligence and Humanized Computing,* vol. 14, no. 8, pp. 10417-10429, 2023, doi: 10.1007/s12652-022-03698-z.

[14] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A pre-trained financial language representation model for financial text mining," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, 2020, doi: 10.24963/ijcai.2020/622.