Stock Price Prediction Based on Investor Sentiment Using BERT and Transformer Models

Chien-Cheng Lee and ANISH SAH
Department of Electrical Engineering, Yuan Ze University
Taoyuan, Taiwan
e-mail: cclee@saturn.yzu.edu.tw, anishkb009@gmail.com

***Abstract*—This paper investigates the impact of investor sentiment on the stock market by predicting stock closing prices and future trends in stock returns. Our study involves gathering abundant investor messages from three social media platforms: Stocktwits, Yahoo Finance, and Reddit. To gauge investor sentiment from the collected messages, we employ Bidirectional Encoder Representations from Transformers (BERT), a transformer-based pre-trained language model. We present a novel application of a Transformer-based model for stock trend prediction. This model architecture leverages the self-attention mechanism to capture the interdependence of stock data, facilitating accurate forecasting of stock trends. By integrating investor sentiment with stock prices and inputting this combined information into the transformer model, we predict the performance of APPLE and SPY stocks datasets. Our experimental results reveal that the transformer model exhibits strong performance regardless of whether sentiment features are included. Moreover, incorporating sentiment does enhance the forecasting accuracy for both stock closing prices and future trends in stock returns.**

Keywords-Bert; Transformer; StockTwits; Stock Returns.

I. INTRODUCTION

In recent years, there has been a growing trend in utilizing text mining technology to automatically extract and analyze substantial amounts of textual data. By employing Natural Language Processing (NLP) techniques, these opinions are summarized, and their applications extend to various domains, including market forecasting [1]. Prior research has explored the financial implications of investor sentiment using conventional NLP techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) [2]. Conventional methods like TF-IDF have limitations in effectively capturing the overall sentiment of complete sentences due to their reliance on lexical frequency analysis. Fortunately, the advancements in NLP have led to the emergence of Bidirectional Encoder Representations from Transformers (BERT), a groundbreaking technology for language modeling [3]. BERT offers robust semantic representation and has demonstrated remarkable performance across various natural language understanding tasks. Its capabilities address the challenges faced by traditional approaches and present new opportunities for more accurate sentiment analysis in financial applications.

However, BERT models are pretrained on diverse text sources like Wikipedia and BookCorpus datasets, which differ considerably from the language used in stock market and economic narratives. This can result in challenges when directly applying these pretrained models for investor sentiment prediction, particularly in interpreting technical terms specific to the financial domain. To address this issue, specialized stock market BERT models are necessary.

The Transformer model, initially introduced by Vaswani et al. [4] in 2017, has brought about a revolutionary change in the realm of deep learning, especially in NLP tasks. However, its significance goes beyond NLP, as researchers have increasingly acknowledged its potential for addressing diverse problems, including stock prediction [5]. The strength of the Transformer model lies in its ability to capture long-range dependencies and learn complex patterns in sequential data. These attributes make it particularly suitable for modeling stock price time series, which are characterized by complex dynamics and influenced by various factors.

In this study, we investigate the influence of investor sentiment on the stock market. To achieve this, we gather investor messages from social media platforms and employ them to pretrain a specialized stock market BERT model for predicting investor sentiment. Subsequently, we conduct experiments by combining investor sentiment with Transformer models to predict stock closing prices and forecast future trends in stock returns. The experimental results reveal that the transformer model exhibits strong performance in both predictions. Furthermore, the incorporation of sentiment features significantly enhances the accuracy of the predictions.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes dataset collection, language model training, sentiment prediction, and stock price and trend prediction. Section IV provides an evaluation of our approach by analyzing the obtained results and presenting comparisons with other methods. Finally, Section V offers some concluding remarks.

II. RELATED WORK

Lexicon-based methods rely on calculating the sentiment score of a vocabulary based on word frequency in optimistic and pessimistic texts. For instance, Oliveira et al. [2] used TF-IDF to determine sentiment scores for the vocabulary. However, the creation of lexicons poses challenges and limitations. One of the primary concerns is that general lexicons may not be well-suited for sentiment analysis in the stock market domain due to certain terms, such as "undervalued," having opposite sentiment interpretations in financial contexts.

 ISBN: 978-1-68558-103-9

In contrast, language model-based methods leverage models like BERT, which have emerged as robust language representation models pretrained on extensive amounts of unlabeled text. Howard and Ruder [7] have proposed various fine-tuning approaches for BERT, including universal language model fine-tuning (ULMFiT).

After conducting sentiment analysis, multiple studies have explored the impact of investor sentiment on the stock market. Kim et al. [8] utilized statistical methods and Naïve Bayes classification to determine investor sentiment from Yahoo Finance messages, evaluating its predictive power for stock returns. Meanwhile, Renault [6] derived investor sentiment from Stocktwits messages and demonstrated its efficacy in forecasting using linear regression models.

Deep learning techniques have also gained substantial traction in stock market research. For instance, Zhong and Enke [9] developed a neural network model that employs economic-related features to predict daily stock market return directions. Wang et al. [10] explored the use of a Transformer model for predicting stock market indices. By incorporating a self-attention mechanism to capture complex relationships in stock market data, the model exhibits higher accuracy than traditional forecasting methods. Additionally, Zhang et al. [11] introduced the Transformer Encoder-based Attention Network (TEANet) framework. This approach effectively captures temporal dependencies and facilitates accurate analysis of financial data. The application of these deep learning techniques marks significant progress in stock market prediction and analysis.

III. MATERIALS AND METHODS

A. Investor Message Dataset Collection

We designed a Python web scraping program to gather investor messages from three prominent social media platforms: Stocktwits, Yahoo Finance, and Reddit (specifically, the World News and News communities). Our data collection from Stocktwits spanned from July 2009 to December 2021. To ensure data integrity, we eliminated duplicate messages and messages containing solely URL addresses and emojis. Ultimately, we retained approximately 34 million messages. Among these, around 13 million were marked with sentiment labels: 11 million were labeled as bullish, 2 million as bearish, while the remaining 21 million were unmarked.

For Yahoo Finance, we collected data from the years 2016 to 2019, which amounted to approximately 1.4 million messages. The Reddit data was gathered during the period from July 2020 to November 2021, amounting to around 60,000 messages. The limited amount of data is due to the difficulty of capturing data, most of which is unlabeled.

B. Stock Market BERT Language Model

For our language model, we selected the BERT-Based pre-trained model provided by Google, which leverages general domain text and two advanced training techniques: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In this study, we took the pre-trained BERT model and performed additional pre-training, specifically tailored to the stock market domain. We accomplished this by utilizing unlabeled data from Stocktwits, Yahoo Finance, and Reddit, employing the MLM approach. This further pre-training of the model allows it to better understand the intricacies and nuances of the stock market language, enhancing its effectiveness in predicting investor sentiment for stock-related tasks.

C. Sentiment Predict Model

MLM further pre-trained model is used to build investor sentiment predictor by fine-tuning the target task classifier with binary output, bullish and bearish. The sentiment predictor uses the BertForSequenceClassification model implemented by the Hugging face library, with a single linear layer added on top for classification. In this study, the accuracy of the sentiment classification model on our validation dataset was 89%. Examples of sentiment prediction results are shown in Table I.

TABLE I. EXAMPLES OF SENTIMENT PREDICTION RESULTS FOR STOCKTWITS

Examples message	Label	Prediction
$SPY 2nd break of the trendline	0	0
$aapl liquidation of options friday hit this stock. a massive winning move!	1	1
$aapl rising corporate debt for stock buybacks. financial engineering vs innovation.	0	0
$FB Careful here. Speculation is overwhelming as shown in this chart.	1	1

After constructing the investor sentiment predictor, we can utilize it to predict whether each unlabeled message exhibits bullish or bearish sentiment. Subsequently, we calculate the daily sentiment index (S_t) using the following formula (1):

$$S_t = \frac{O_t - P_t}{N_t} \quad (1)$$

where O_t is the number of bullish messages on day t, P_t is the number of bearish messages on day t, and N_t is the total number of messages on day t. In this way, the daily sentiment index S_t ranges from -1 to 1. A negative value indicates prevailing bearish sentiment, a positive value indicates prevailing bullish sentiment.

D. Transforme Learning Models for the Impact of Investor Sentiment

In this study, we use the Transformer model to predict stock trends and evaluate its effectiveness in capturing the intricate dependencies and patterns within our stock dataset. The Transformer encoder combines self-attention and feed-forward layers, enabling it to efficiently capture the relationship between tokens in a sequence. By leveraging these mechanisms, our model can learn and exploit the dependencies among different stock data elements.

The subsequent layers of our model consist of global average pooling, dense layers, and softmax activation, collectively responsible for generating the classification output. Global average pooling is used to condense the sequence into a fixed-length representation for further analysis and decision-making. Dense layers introduce non-

linearity and higher-level representations, that enhance the model's ability to understand complex patterns. Finally, a softmax activation function predicts stock price returns by assigning probabilities to different classes.

During our experiments, we explored various parameter settings. The best-performing configuration utilizes one Transformer encoder block, one multi-head attention layer, and two convolutional layers used as a feed-forward network. To prevent overfitting and improve generalization, we add dropout after the multi-head attention layer. The training process of our model involves using the Adam optimizer and the sparse categorical cross-entropy loss function. The model's performance was assessed using the sparse categorical accuracy metric, measuring the accuracy of the predicted stock returns.

IV. Experimental Results

To study the impact of investor sentiment on the stock market, we conducted two experiments: predicting stock closing prices and predicting future trends in stock returns. The experiments were performed on a computer with an NVIDIA Geforce GTX 1080Ti GPU card with 32 GB of memory. We used Tensorflow to implement the models. For computational reason, we only investigated Apple (ticker: AAPL) and S&P 500 ETF (ticker: SPY). Stock price data, including daily opening, highest, lowest, adjusted closing prices, and closing prices, was downloaded from July 2010 to November 2021 on Yahoo Finance. Investor messages from Stocktwits over the same time period for these two stocks were also collected. The total number of sentiment messages for AAPL and SPY were 530,099 and 1,823,709, respectively.

In our analysis, we designated the data from July 2010 to December 2019 as the training data, while the data after that period was assigned as the test data. As part of the normalization process, we scaled each individual data column, particularly the stock price data, to fall within the range of [0, 1]. After prediction, we restore the output to its original range using the stored normalization parameter.

A. Prediction Results of Closing Prices

To demonstrate the impact of sentiment, we create two feature sets: one comprising both sentiment index and stock price data (daily opening, highest, lowest, and closing prices), and the other containing only stock price data. We utilize a 25-day data window to forecast one day ahead, specifically predicting the data for the 26th day. Subsequently, the feature set is input into the Transformer models to make predictions on stock closing prices.

In order to evaluate the prediction performance, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-square (R^2) are used as evaluation criteria in this study. Their formulas are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y_i}| \quad (2)$$

$$RMSE = \sqrt{\frac{i}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Here, n is the total number of samples; y_i and $\hat{y}_i$ represent the true value and predicted value of the test set, respectively. The smaller the *MAE* value, the better the prediction. Likewise, the smaller the *RMSE* value, the better the prediction. That is, the closer the values of *MAE* and *RMSE* are to 0, the smaller the error between the predicted value and the true value. In R^2, the mean of the true values of the test set is not directly represented. Instead, R^2 measures the proportion of the variance in the dependent variable (target) that is explained by the independent variables (predictors) in the model. The value range of R^2 is (0, 1). The closer R^2 is to 1, the better the model fits the data.

Table II shows the closing price prediction results of the test data (including MAE, RMSE and R2 standards) of two stocks compared with Long Short-Term Memory (LSTM). The inclusion of sentiment features in the predictions results in better performance compared to predictions without sentiment features. Furthermore, the presence of sentiment features leads to an improvement in the accuracy of closing price prediction. Figure 1 illustrates the superior closing price prediction curves of the Transformer model for two stocks when using 25-day data with sentiment features. Our predicted closing price curves closely align with the true price curves, demonstrating the model's remarkable accuracy in capturing the stock price trends.

B. Prediction Results of Future Trends in Stock Returns

Most investors are concerned with future trends in stock returns rather than the actual price value. Hence, we adopt accuracy as a measure to assess the prediction performance of future trends in stock returns. The stock return is determined based on whether the closing price is up or down, and any changes in the closing price will consequently alter the stock return. The stock return $R_{t,i}$ of stock i on date t is calculated as follows:

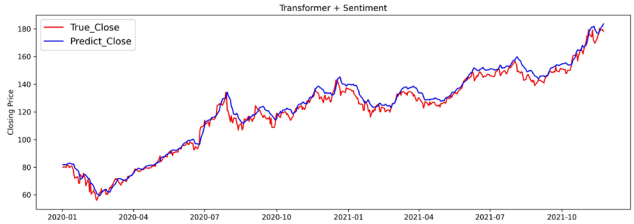
$$R_{t,i} = \frac{Close_i(t) - Close_i(t-1)}{Close_i(t-1)} \times 100 \quad (5)$$

where $Close_i(t)$ is the closing price of stock i on date t. A positive return indicates that the closing price today is higher than the closing price of the previous day. In classification problems, accuracy is a commonly used evaluation metric. To convert the evaluation of stock return regression into a binary classification, we define the labels for future trends as follows: a label of 1 represents an expected increase in stock return, while a label of 0 indicates a decrease or no change in stock return. By adopting this approach, we can assess the model's performance in predicting the direction of future stock returns, which is more intuitive for investors seeking to understand potential trends.

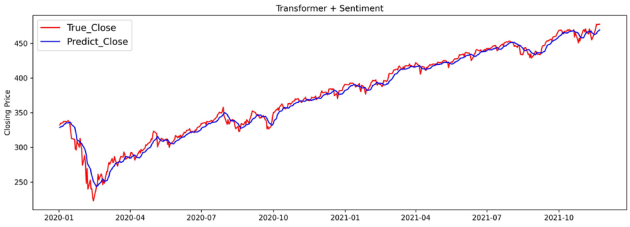
First, we calculate the daily stock return, denoted as $R_{t,i}$, and integrate it into the existing stock price data. This updated dataset includes the daily open, high, low, close prices, along with the corresponding stock returns. Next, we employ Transformer models to build stock return prediction models, allowing us to evaluate and compare their performance. By

 ISBN: 978-1-68558-103-9

incorporating stock returns, we aim to enhance the accuracy and effectiveness of our predictions, leading to valuable insights for investment decision-making.



(a)



(b)

Figure 1. Closing price prediction curves of the Transformer model for stocks using 25-day data with sentiment features. (a) AAPL, (b) SPY

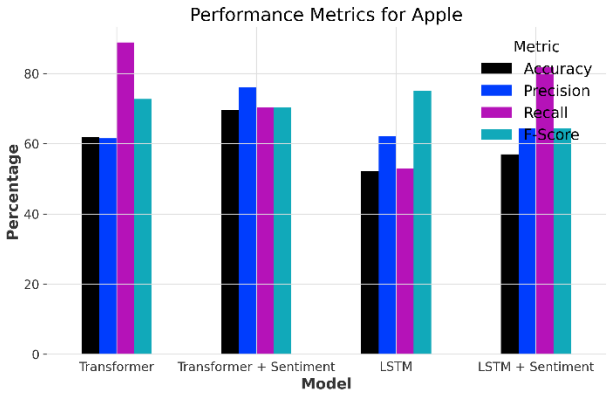
TABLE II. PREDICTION RESULTS OF CLOSING PRICES

Stock	Models	25- Days Data		
		MAE	*RMSE*	R^2
AAPL	Transformer	0.0301	0.0404	0.9878
	Transformer + Sentiment	0.0279	0.0374	0.9882
	LSTM	0.0372	0.0483	0.9829
	LSTM+ Sentiment	0.03298	0.0429	0.9823
SPY	Transformer	0.0176	0.0264	0.9885
	Transformer + Sentiment	0.0169	0.0217	0.9855
	LSTM	0.0271	0.0328	0.9825
	LSTM + Sentiment	0.0194	0.0230	0.9837

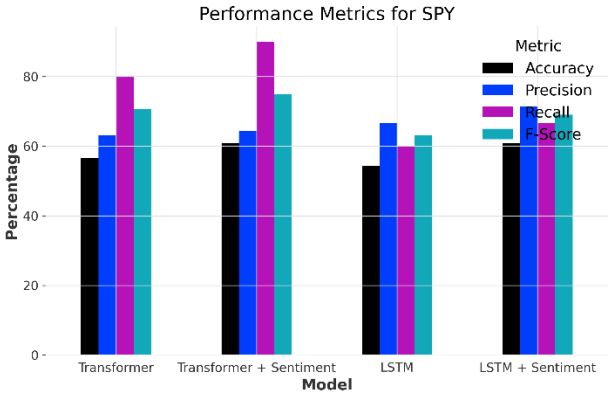
TABLE III. PREDICTION METRICS OF FUTURE TRENDS IN STOCK RETURNS (IN %)

Model	*Stock*	*Accuracy*	*Precision*	*Recall*	*F-Score*
Transformer	Apple	61.85	61.53	88.88	72.72
	SPY	56.52	63.15	80.00	70.58
Transformer + Sentiment	Apple	69.56	76.00	70.37	70.37
	SPY	60.86	64.28	90.00	75.00
LSTM	Apple	52.18	62.06	52.94	75.00
	SPY	54.34	66.67	60.00	63.15
LSTM + Sentiment	Apple	57.00	64.28	81.81	64.28
	SPY	60.86	71.42	66.66	68.96

To evaluate the influence of sentiment, we create two sets of features: one with sentiment and the other without sentiment. Table III and Figure 2 show the performance metrics compared to LSTM for predicting future trends in stock returns using 25 days of data. incorporating sentiment as a feature significantly improves the accuracy of these predictions. Notably, AAPL stock demonstrates the highest accuracy, achieving an impressive 70%. These findings highlight the valuable impact of sentiment in enhancing the precision of stock return predictions, particularly for the AAPL stock.



(a)



(b)

Figure 2. Visualized performance metrics for the stock data of (a) AAPL and (b) SPY.

V. CONCLUSIONS

In this paper, we employ BERT for sentiment classification in the stock market, focusing on individual stocks. Following this, we conduct a series of experiments to investigate the impact of investor sentiment on the stock markets. By harnessing the capabilities of the powerful Transformer model and optimizing its parameters, our experimental results reveal that incorporating sentiment information can potentially offer substantial benefits in enhancing the accuracy and effectiveness of stock market predictions. The findings underscore the value of sentiment analysis in better understanding market dynamics and making more informed investment decisions.

REFERENCES

[1] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, “Text mining for market prediction: A systematic review,”

Expert Systems with Applications, vol.41, no. 16, pp. 7653-7670, 2014.

[2] N. Oliveira, P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures," Decision Support Systems, vol. 85, pp. 62-73, 2016.

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

[4] A. Vaswani, et al., "Attention is all you need," The 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, 2017.

[5] T. Muhammad, et al., "Transformer-Based Deep Learning Model for Stock Price Prediction: A Case Study on Bangladesh Stock Market," ArXiv, 2022. abs/2208.08300.

[6] T. Renault, "Intraday online investor sentiment and return patterns in the US stock market," *Journal of Banking & Finance*, vol. 84, pp. 25-40, 2017.

[7] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146, 2018.

[8] S. H. Kim and D. Kim, "Investor sentiment from internet message postings and the predictability of stock returns," *Journal of Economic Behavior & Organization*, vol. 107, pp. 708-729, 2014.

[9] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," Financial Innovation, vol. 5, no. 1, pp. 1-20, 2019.

[10] C. Wang, Y. Chen, S. Zhang, and Q. Zhang, "Stock market index prediction using deep Transformer model," *Expert Systems with Applications*, vol. 208, 118128, 2022.

[11] Q. Zhang, et al., "Transformer-based attention network for stock movement prediction," *Expert Systems with Applications*, vol. 202, 15, 2022.

 ISBN: 978-1-68558-103-9