

# 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](mailto:cclee@saturn.yzu.edu.tw)

# My Bio

- **Chien-Cheng Lee** (Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2003.
- I'm currently an Associate Professor with the Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan.
- I was a Visiting Researcher with Telcordia Inc. (formerly, Bellcore), Piscataway, NJ, USA, from October 2007 to January 2008.
- My research interests include image processing, pattern recognition, natural language processing (NLP), and machine learning.

# Introduction

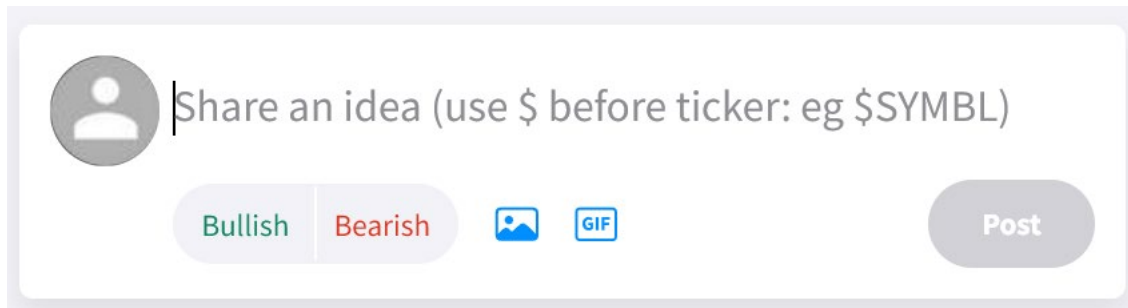
- Social media has grown rapidly over the past 15 years, such as Twitter, Facebook, and Reddit.
- People view, judge and provide their opinions online, which may contain valuable information.
- This valuable information may influence the decision-making process.
- Therefore, in recent years, many studies have utilized natural language processing (NLP) technology to automatically mine and analyze large amounts of text.

# Introduction

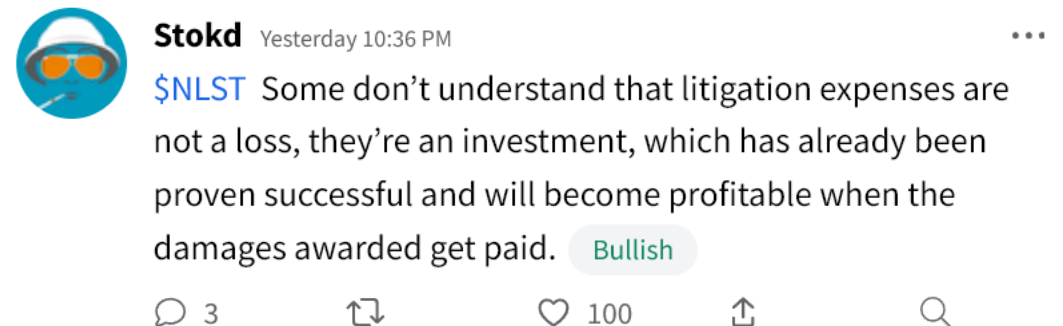
- This paper studies the impact of investor sentiment on the stock market by predicting stock closing prices and future trends in stock returns.
- We gather extensive investor messages from three social media platforms: Stocktwits, Yahoo Finance, and Reddit.
- The NLP technology, BERT (Bidirectional Encoder Representations from Transformers) language model, is used to estimate the investor sentiment from the collected investor messages.
- We combine investor sentiment with stock price data and feed it into a Transformer model to predict the closing stock prices of APPLE and SPY stocks as well as future trends in stock returns.

# Investor Message Dataset Collection

- We developed a Python web scraping program to collect investor messages from three social media platforms: Stocktwits, Yahoo Finance, and Reddit (World News and News communities).
- Stocktwits is the largest investor and trader community website in the U.S. stock market.
- In September 2012, Stocktwits added a new feature that allows users to directly express their opinions and mark them as **bullish** or **bearish**.



A screenshot of the Stocktwits post creation interface. It features a text input field with a placeholder that says "Share an idea (use \$ before ticker: eg \$SYMBL)". Below the input field are two buttons labeled "Bullish" and "Bearish", followed by icons for adding images and GIFs. A "Post" button is located on the right side of the interface.



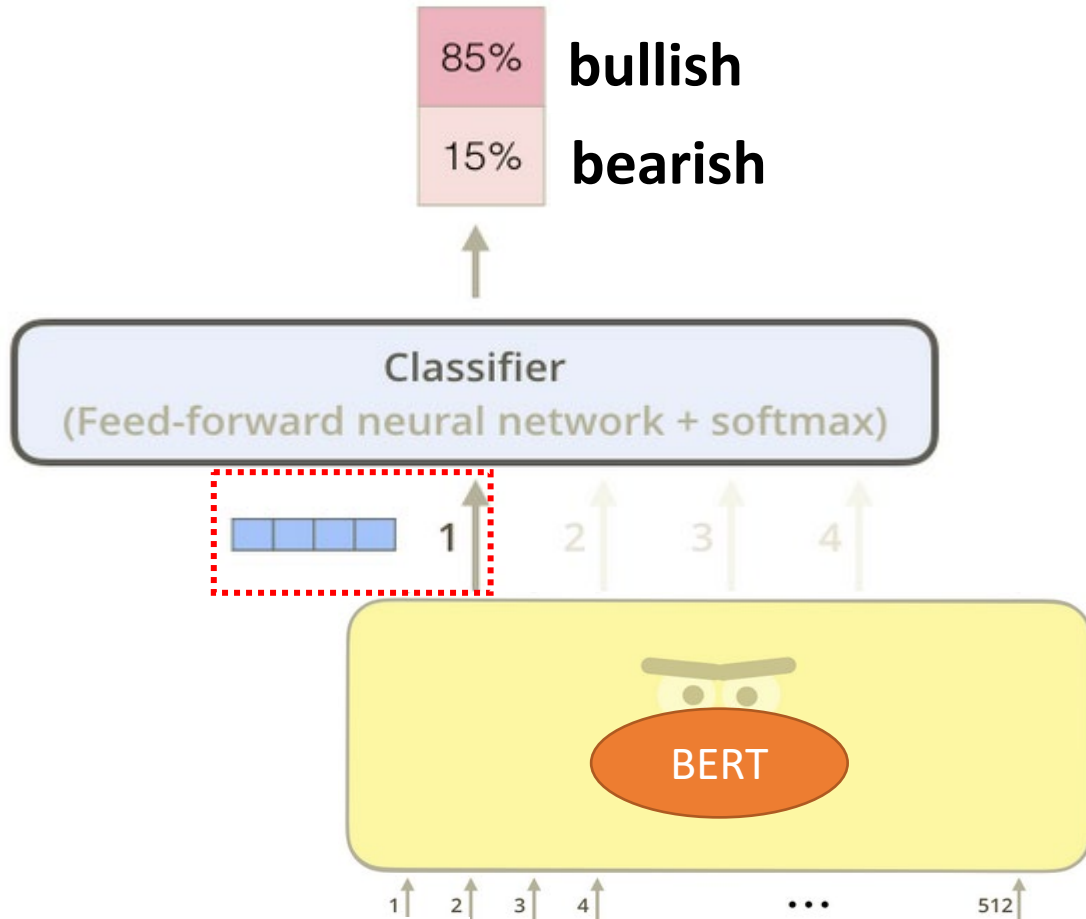
# Dataset size

- Stocktwits
  - We collected 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.
- Yahoo Finance
  - We collected approximately 1.4 million messages.
- Reddit
  - We collected approximately 60,000 messages.

# Stock Market BERT Language Model

- We choose the BERT-based pre-trained model provided by Google as the language model to estimate message sentiment.
- In this study, we take the pre-trained BERT model and performed additional pre-training to transfer the model from the general domain to the stock market domain.
- This further pre-training of the model enables it to better understand the complexities and nuances of the stock market language, enhancing its effectiveness in estimating investor sentiment for stock-related tasks.
- The sentiment predictor uses the BertForSequenceClassification model implemented by the Huggingface library. In this study, the accuracy of the sentiment estimation model on our validation dataset was 89%.

# Sentiment Predictor



Label	Training set	Validation set	Total
Bullish	1,765,426	441,356	2,206,782
Bearish	1,535,153	383,788	1,918,941
Total	3,300,579	825,144	4,125,723

The bearish data is much smaller than the bullish data in the Stocktwits dataset. To keep the training data balanced, we use all bearish messages and randomly select bullish messages at 1.15 times the amount of bearish.

using the first 126 words of the message

[CLS] [\\$AAPL](#) has a Gross Margin of 43.45%. This is in the better half of the industry: [\\$AAPL](#) outperforms 70.59% of its industry peers.

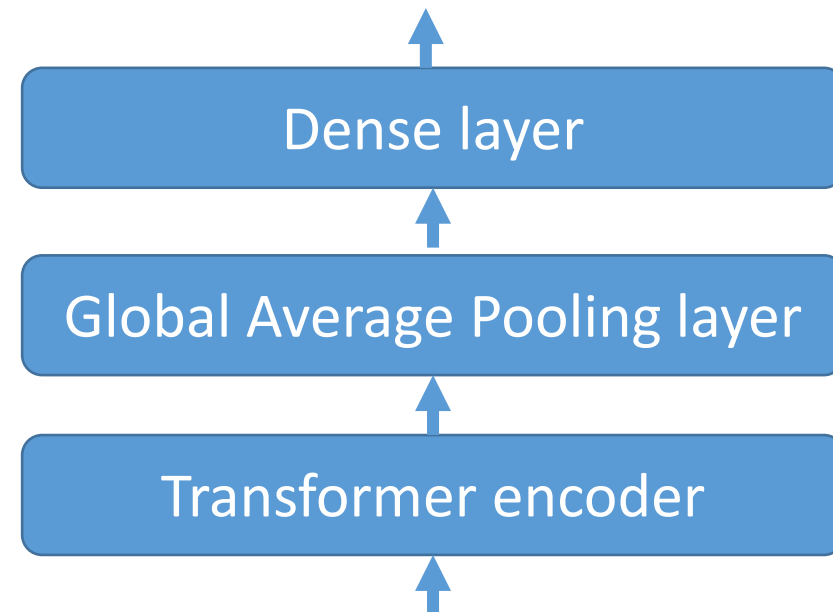


# Transformer Prediction Models

- In this study, we use the Transformer encoder to predict stock closing prices and future trends in stock returns.
  - The core idea behind the Transformer model is *self-attention*—the ability to attend to **different positions** of the input sequence to compute a representation of that sequence.
  - A transformer model handles variable-sized input using stacks of self-attention layers instead of RNNs or CNNs.

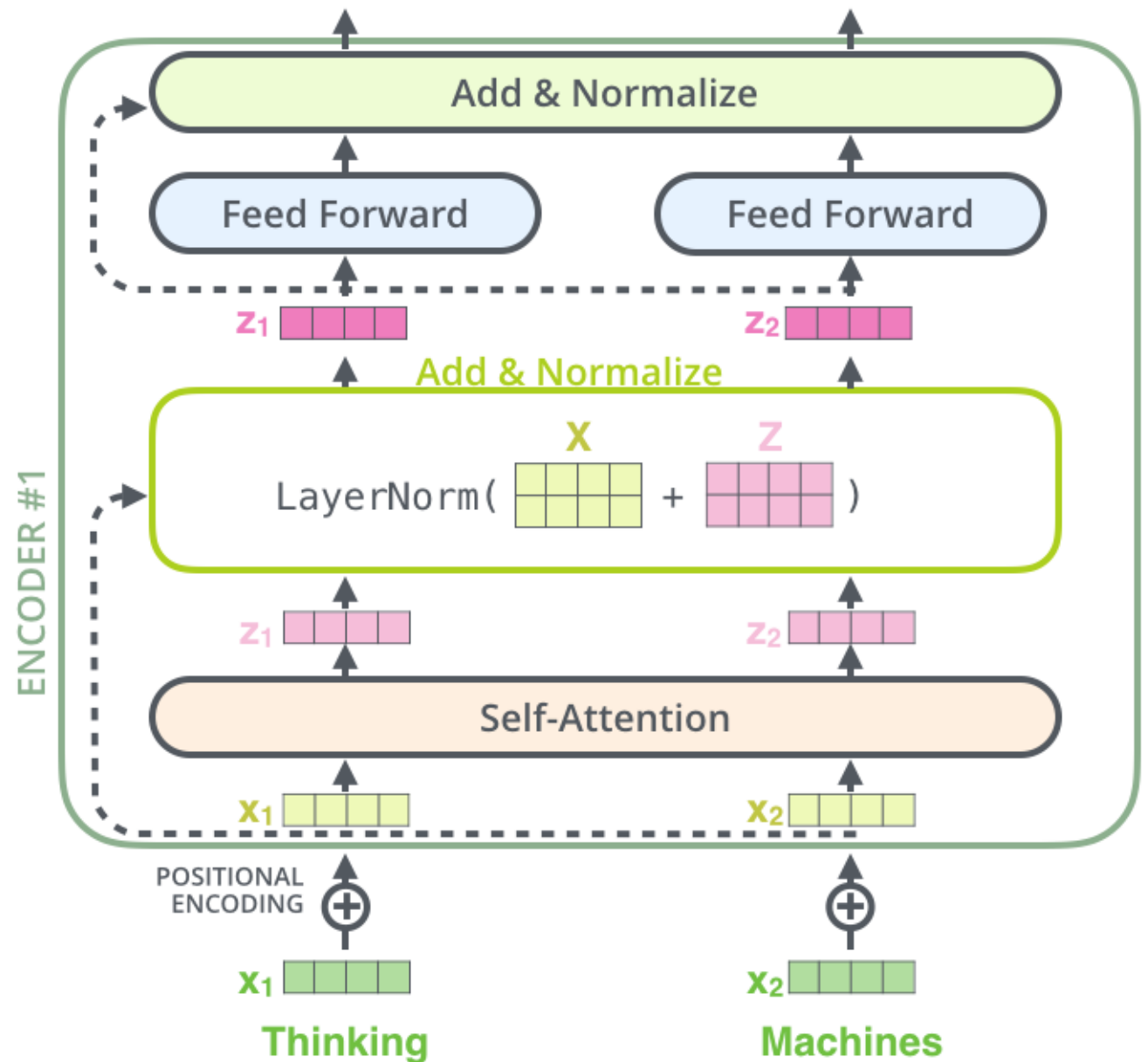
Global average pooling is used to condense the sequence into a fixed-length representation.

Dense layers introduce non-linearity and higher-level representations.



# Transformer Encoder

- Transformer encoder combines self-attention and feed-forward layers, enabling it to efficiently capture the relationship between tokens in a sequence.

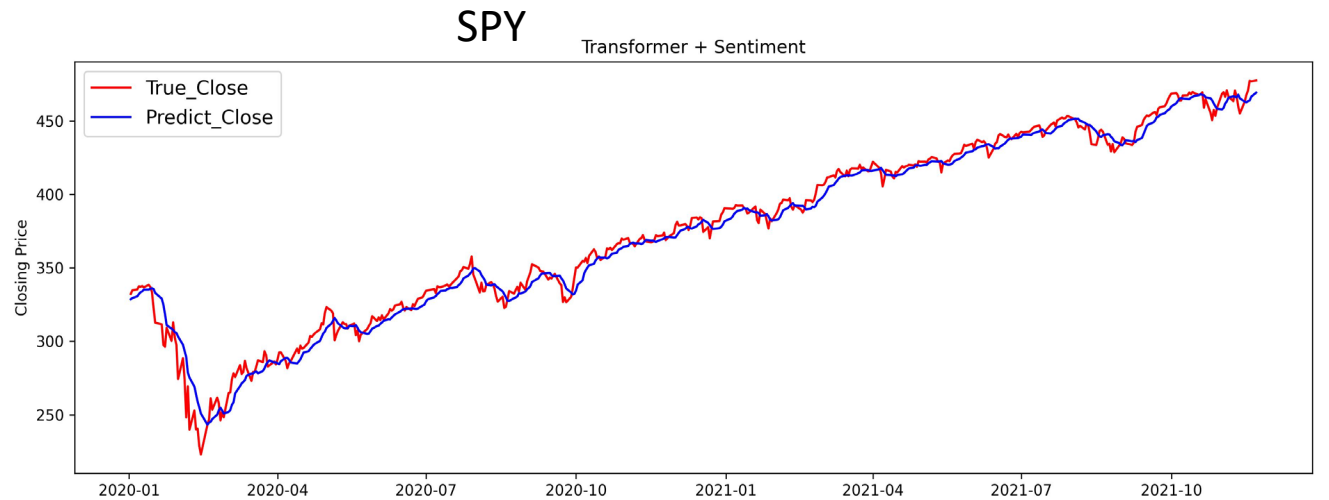
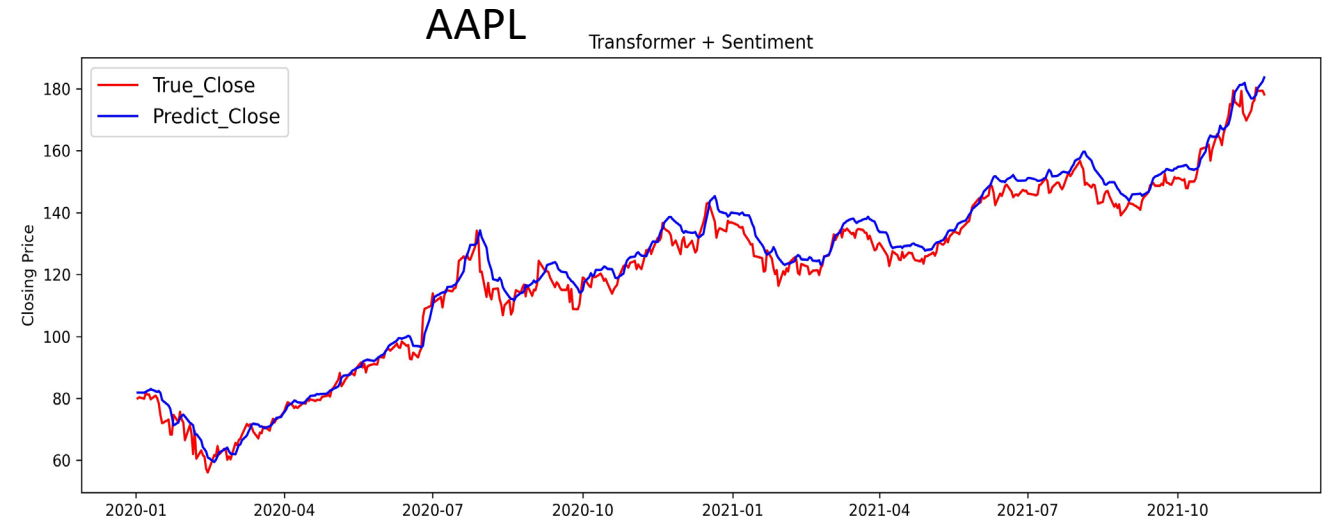


# Experimental Results

- The experiments were performed on a computer with an NVIDIA 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.
- The total number of sentiment messages
  - for AAPL were 530,099
  - for SPY were 1,823,709

# Prediction Results of Closing Prices

- We utilize a 25-day data window to forecast one day ahead, specifically predicting the data for the 26th day.
- Red line: real price
- Blue line: predicted price



# Prediction Results of Closing Prices

- Evaluation criteria

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- Root Mean Square Error (RMSE)

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

- R-square ( $R^2$ )

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

# Prediction Results of Closing Prices

Stock	Models	25- Days Data		
		MAE	RMSE	R <sup>2</sup>
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

# Prediction Results of Future Trends in Stock Returns

- The stock return  $R_{t,i}$  of stock  $i$  on date  $t$  is calculated as follows:

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

- where  $Close_i(t)$  is the closing price of stock  $i$  on date  $t$ .
- 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.
- Now, it's a classification problem.

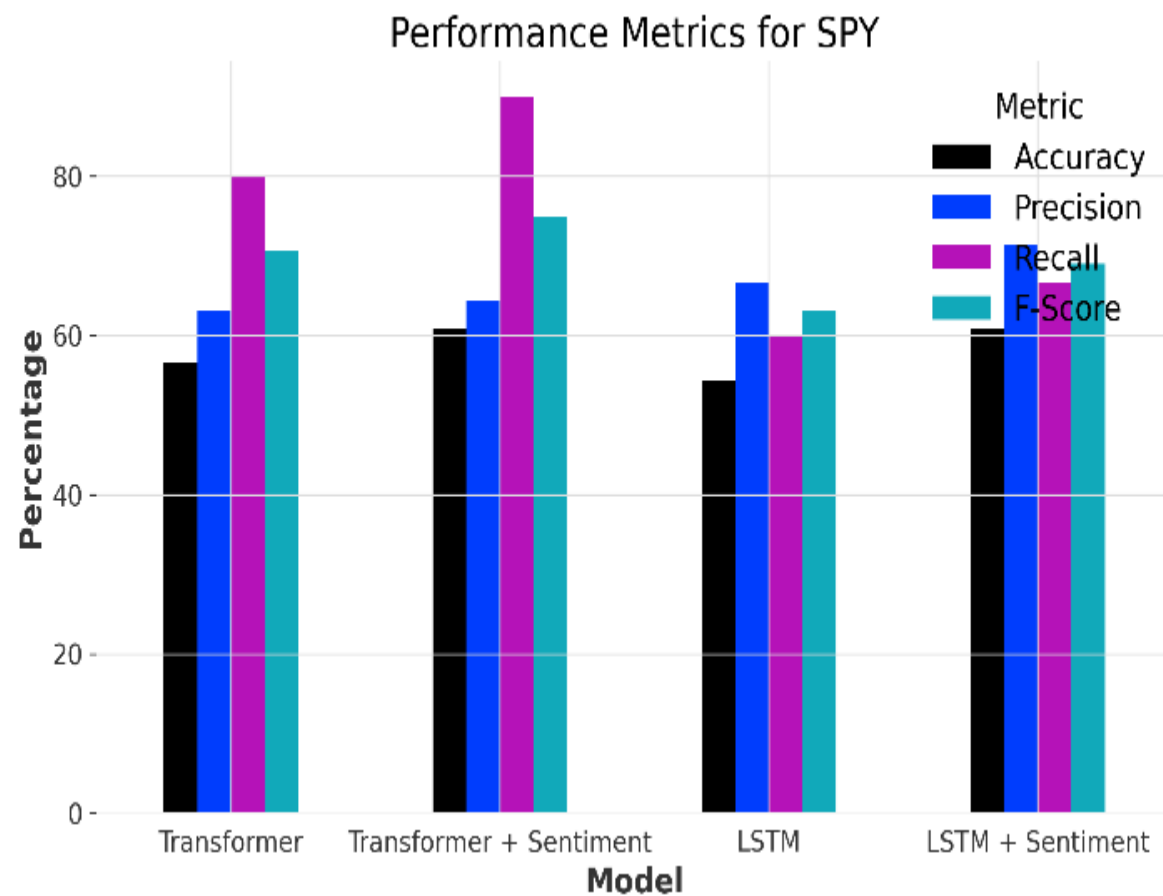
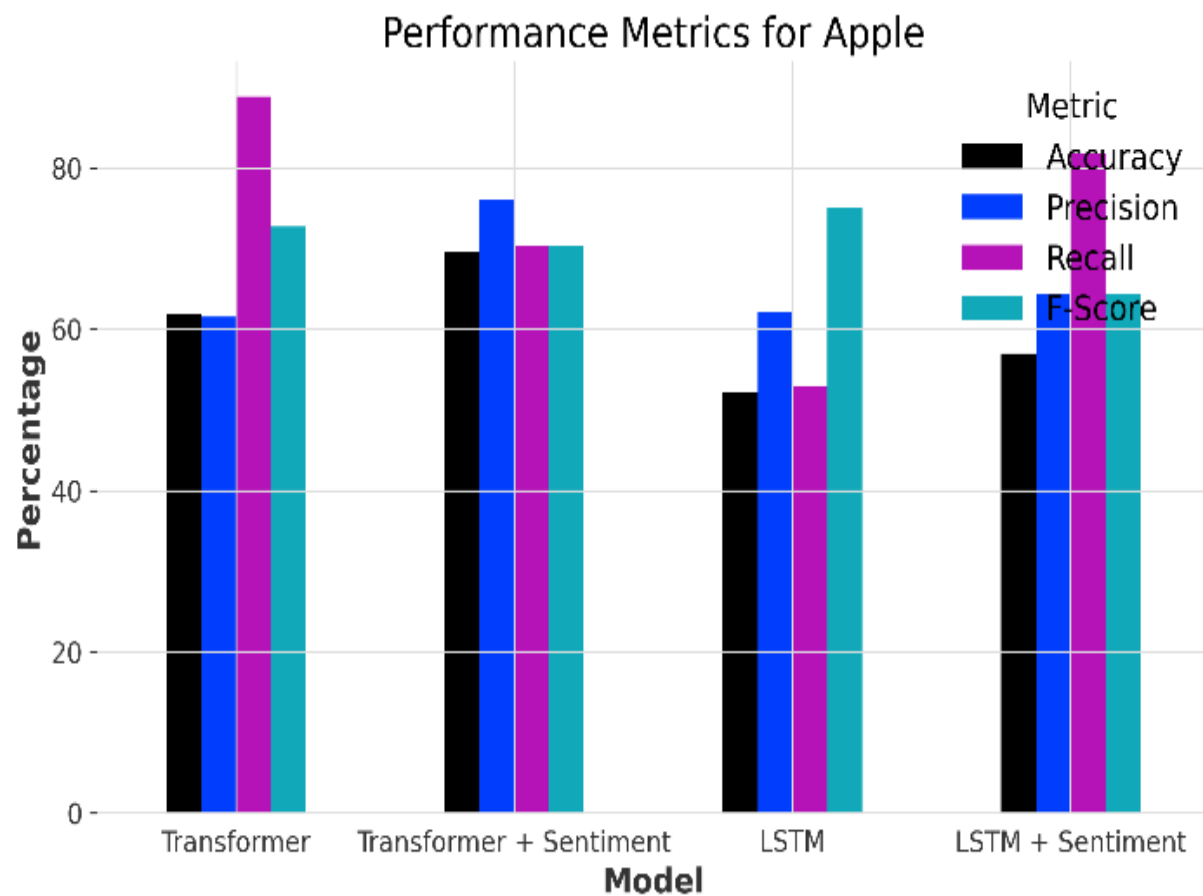
# Prediction Results of Future Trends in Stock Returns

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

The F-score is a way of combining the precision and recall of the model.



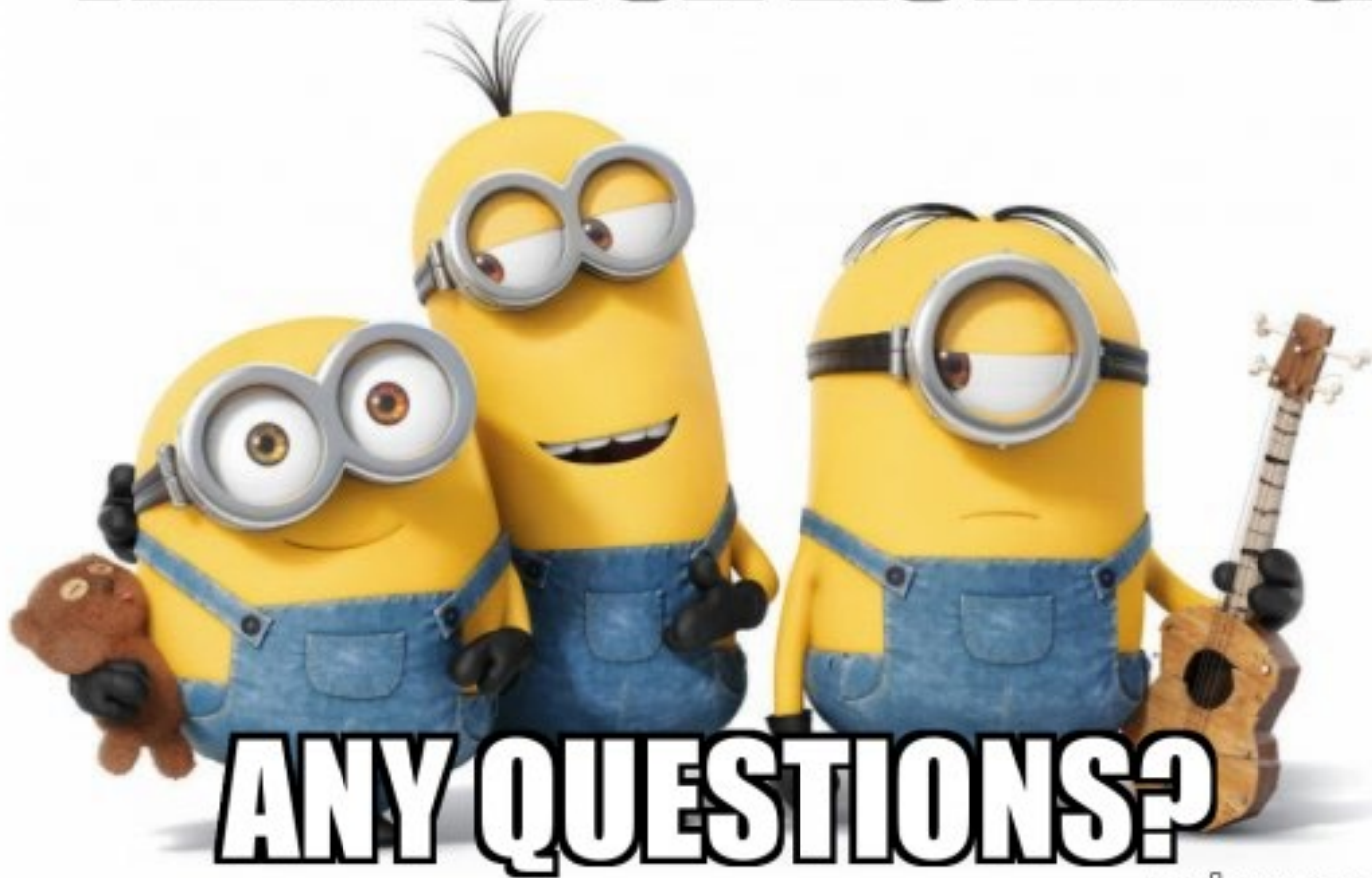
# Prediction Results of Future Trends in Stock Returns



# Conclusions

- In this paper, we employ BERT for sentiment classification in the stock market.
- We conduct a series of experiments to investigate the impact of investor sentiment on the stock markets.
- Our experimental results reveal that incorporating sentiment information can potentially offer substantial benefits in enhancing the accuracy and effectiveness of stock market predictions.

**THANKS FOR LISTENING**



**ANY QUESTIONS?**

[makeameme.org](http://makeameme.org)