RAG-Augmented LLMs: Advancing Stock Predictions Through Data-Driven Retrieval

By Harshit Tomar
(2201AI15)
Under - Dr.Jimson Mathew

CONTENT

Objective:

To improve stock price prediction accuracy by combining statistical models with company-specific insights using LLMs and RAG for semantic understanding and embedding creation.

- RAG can retrieve historical market data and news, enabling the model to understand the context of current market conditions, similar past events, or patterns that could affect stock prices.
- LLMs then synthesize this information to generate predictions considering historical and real-time factors.

APPROACH :

We tried different models and in each model, we use lookback = 30 days (previous 30 days' data to predict the next day for multi-step we predict the next 7 days, LLM based rag system we predict % change in stock price) and gave an average RMSE of 50 different stocks as a common standard to evaluate the models)

LITERATURE SURVEY

Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction: a systematic review. In Journal of Physics: Conference Series (Vol. 2161, No. 1, p. 012065). IOP Publishing.

Pros:

The paper highlights a diverse range of algorithms, including SVM, RNNs, LSTMs, and ARIMA, offering flexibility for practitioners to choose the most suitable method for stock price prediction based on data type and forecasting needs.

Cons:

The paper's analysis is limited to a single model or small dataset, which reduces the ability to generalize findings.

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., & Lin, J. (2024). Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. ArXiv. https://arxiv.org/abs/2402.10350

Pros:

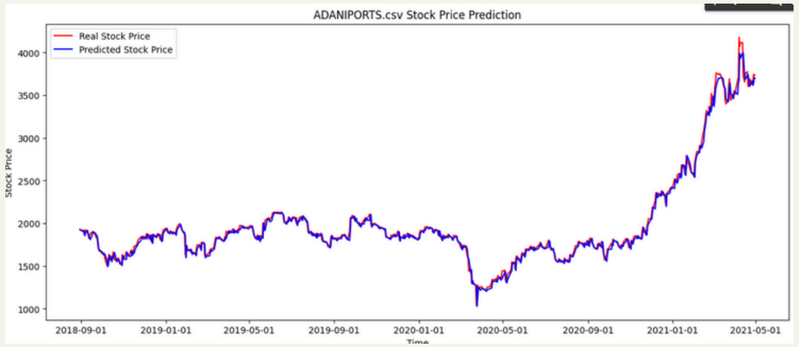
This paper provides a comprehensive systematic literature review on the application of Large Language Models (LLMs) in forecasting, exploring their potential to enhance predictive analytics across various domains.

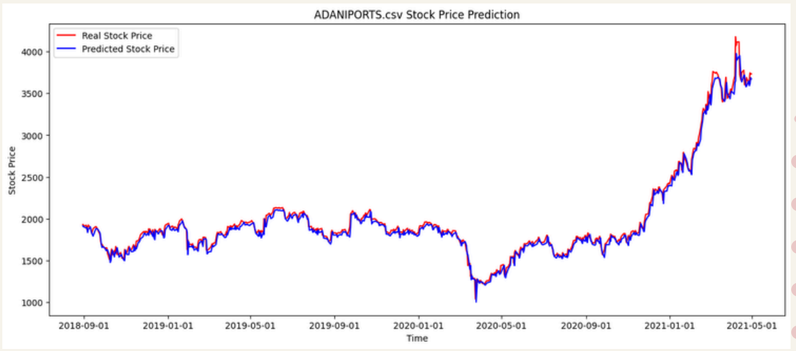
Cons:

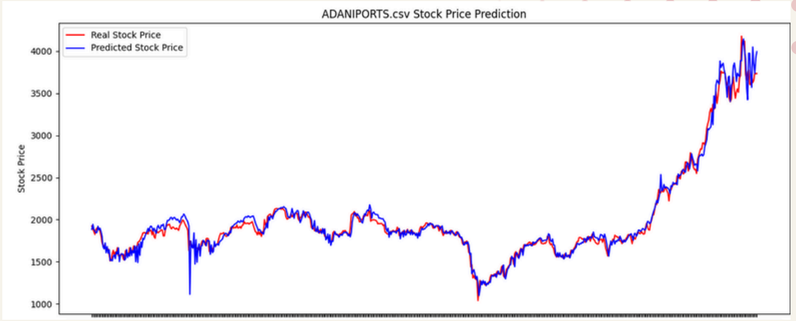
LLMs are not specifically designed for accurate predictions but are more effective in converting text into semantic representations.

RNN ARCHITECTURES

Model	Architecture	No. of Trainable Parameters	Computational Complexity	RMSE (Accuracy)
RNN	Predicts next-day closing price using last 30 days. Hidden Units: 30, Layers: 5, LR: 0.01, Epochs: 100	4,861	$O(H^2 \times T \times L)$	314.79
GRU	Predicts next-day closing price using last 30 days. Hidden Units: 30, Layers: 5, LR: 0.01, Epochs: 100	14,071	$O(3 \times H^2 \times T \times L)$	263.18
LSTM	Predicts next-day closing price using last 30 days. Hidden Units: 30, Layers: 5, LR: 0.01, Epochs: 100	33,151	$O(4 \times H^2 \times T \times L)$	1295.47



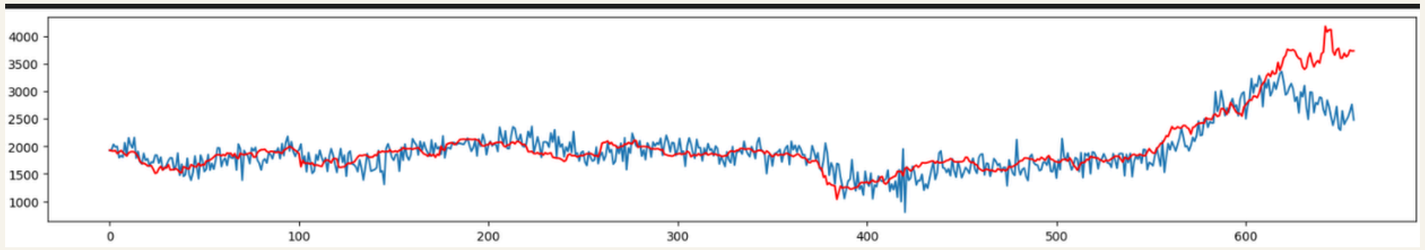




- With increased model complexity (GRU and LSTM), the number of parameters and computational complexity increases.
- with GRU providing a good balance between performance and complexity.

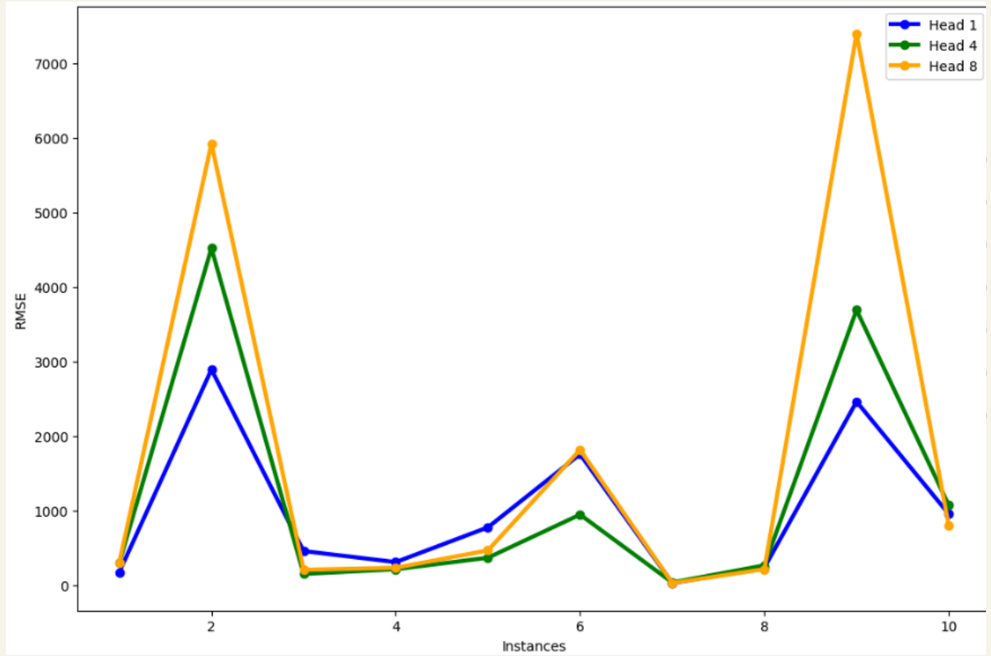
ATTENTION-BASED MODELS

Parameter	Value
Objective	Predicting next-day closing price using the last 30 days
Model Parameters	
- D_model	512
- Number of Heads	8
- Number of Layers	2
- Embedding Size	512
- Dimensionality (dff)	2048
- Trainable Parameters	6,306,305
Training Configuration	
- Learning Rate	0.01
- Epochs	100
Complexity	$O(L \times (T^2 \times D\_model + T \times D\_model \times dff))$
Performance Metric (RMSE)	1182.69



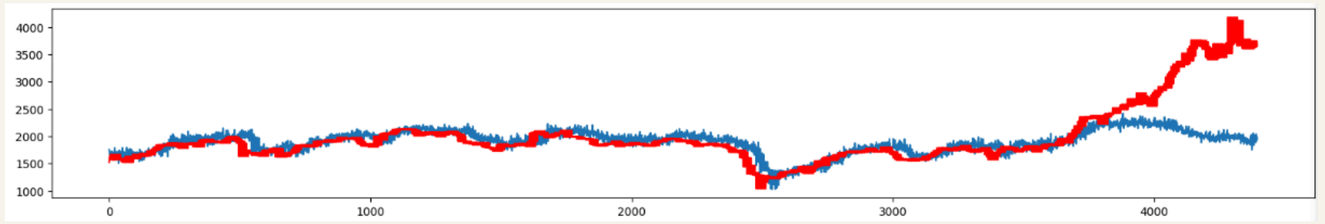
- Single-step prediction using Transformer architecture (encoder with dense layer):
- High RMSE in the case of both low number of attention heads and high number of heads.

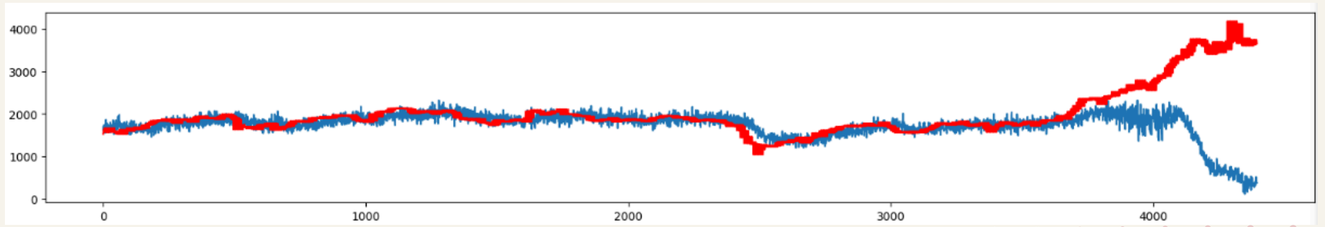
Showing Rmse Variation For Different no of head with keeping the no of taianble parameters same



MULTISTEP PREDICTIONS

Predicting the next 7 days using the last 30 days with two different architectures: one is attention-based Transformer model (encoder with dense layer), and the other is the attention-free FNet.





Feature	Attention-Based Transformer	Attention-Free FNet
Input Shape	`(None, 30, 1)`	`(None, 30, 1)`
Encoder Output Shape	`(None, 30, 512)`	`(None, 30, 512)`
Slicing Operation	`(None, 7, 512)`	`(None, 7, 512)`
Dense Layer Output	`(None, 7, 1)`	`(None, 7, 1)`
Trainable Parameters	6,306,305	4,205,057
RMSE	774.16	468.89

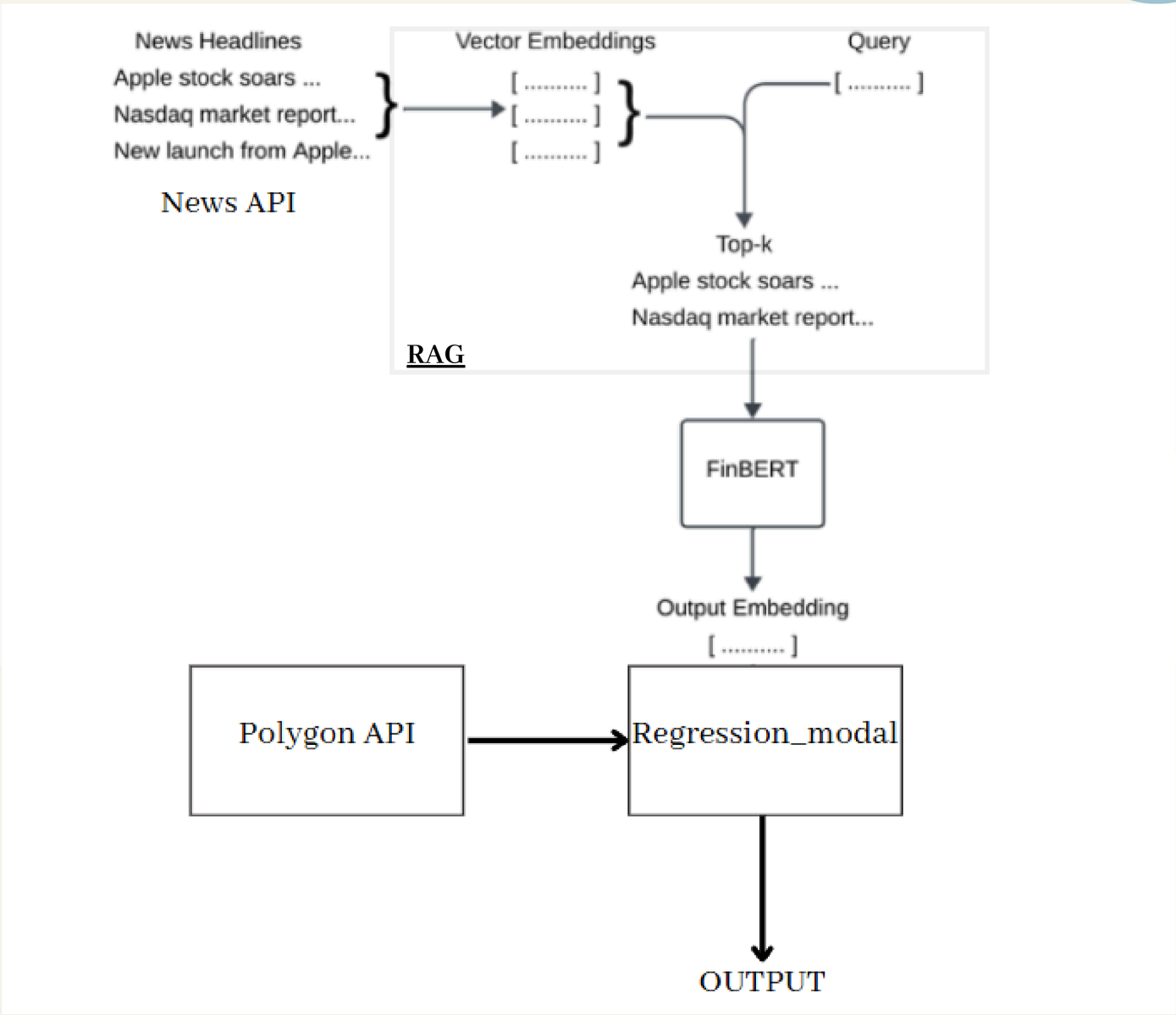
RAG-AUGMENTED LLMS

Output Embedding:
The model generates predictions by passing tokenized text to FinBERT and using the [CLS] token for regression

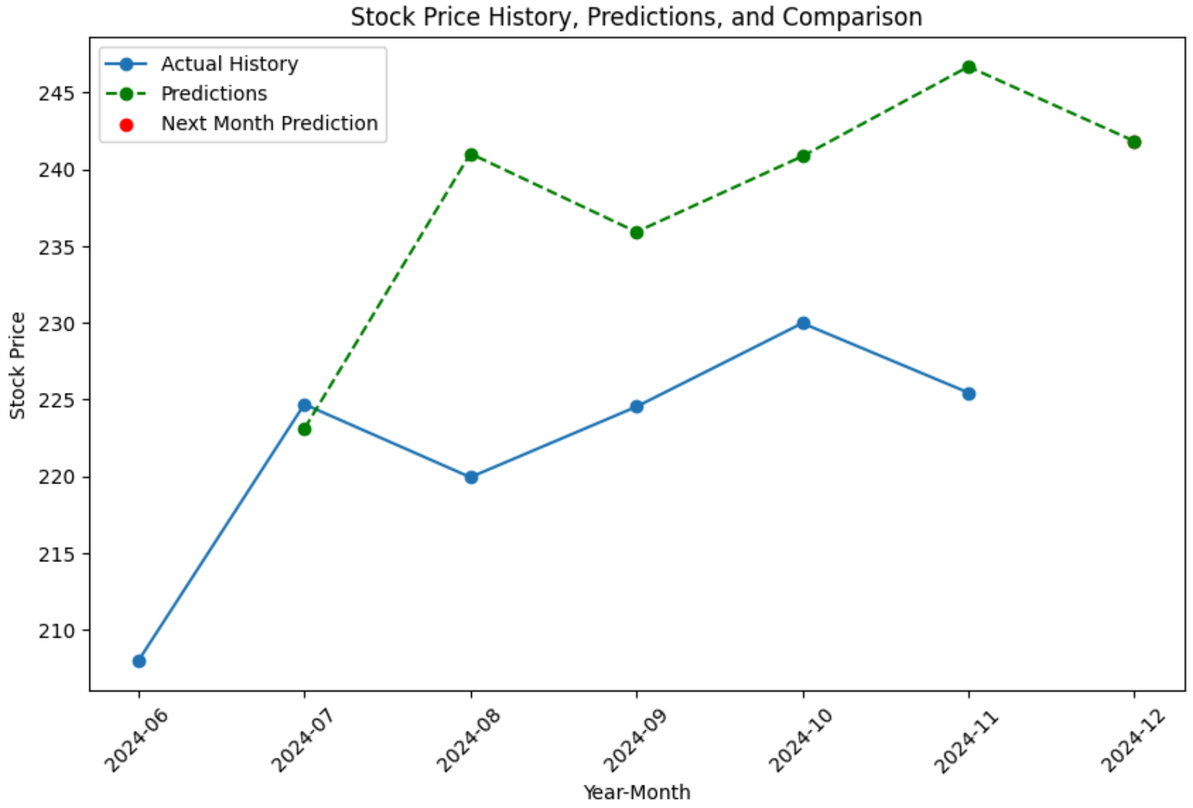
Output:
Percentage Increase or Decrease in the Trend of Stock Price for the Next Month

Regression model:
Linear(in_features=768, out_features=1

Query:
Filter and prioritize news headlines explicitly discussing factors influencing {company_name}'s stock price, including market trends, financial reports, corporate announcements, and industry-specific news impacting {company_name}'s performance.

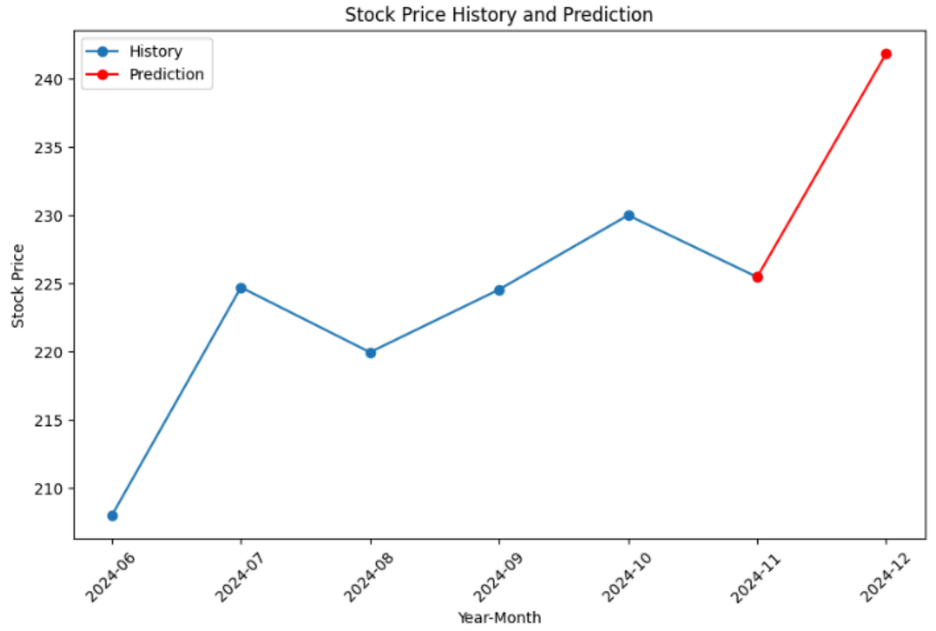


```
Comparison of Predicted vs. Actual Prices for Past Months:
2024-06: Predicted: 223.10, Actual: 224.69, Error: -1.60
2024-07: Predicted: 241.02, Actual: 219.94, Error: 21.09
2024-08: Predicted: 235.92, Actual: 224.53, Error: 11.39
2024-09: Predicted: 240.85, Actual: 229.97, Error: 10.87
2024-10: Predicted: 246.69, Actual: 225.45, Error: 21.24
```




```
FinancialBERT Monthly Stock Price Change Rate:
7.267480343580246% change from current month's stock price (vw)


Polygon Stock Price History:
2024-06 stock price (vw): 207.98
2024-07 stock price (vw): 224.69
2024-08 stock price (vw): 219.94
2024-09 stock price (vw): 224.53
2024-10 stock price (vw): 229.97
2024-11 stock price (vw): 225.45
2024-12 stock price (vw): 241.83 (predicted)
```



THANK YOU

Presented By : Harshit Tomar