Sequential Reinforcement Learning on Economic Discourse for Real-Time Financial Trading

Chirag Maheshwari, Sarang Goel, and Ekansh Mittal

Department of Computer Science, Stanford University

Stanford
Computer Science

Project Overview

Problem Statement

- **Live Discourse:** Financial markets increasingly depend on real-time narrative information: earnings calls, central bank announcements, macroeconomic reports.
- **Interpretation:** Existing algorithmic trading agents rely on structured price data and cannot react to live economic language.
- **Objective:** Build a reinforcement learning (RL) agent that processes financial discourse line-by-line, integrating language and market state to make real-time trading decisions

Significance

- **Narratives:** Market-moving events are often triggered by narrative information, not just price patterns**.** Human traders listen and react to earnings calls and economic announcements in real time.
- **Automation Advantages:** AI systems that can understand and act on narrative information would significantly improve trading responsiveness.
- **Scalability:** As global financial discourse grows, human traders cannot keep pace; adaptive AI agents are needed.}
- **Potential impact:** Narrative-aware trading agents could outperform conventional systems and enable more explainable, human-aligned strategies.

Existing Approaches & Limitations

- **FinRL:** Operates on structured price data; ignores language.
- **FinBERT:** Static sentiment classification; reduces complex financial discourse to oversimplified labels.
- **Deep RL for HFT:** Optimizes microsecond trading on order book data; ignores macro-level narrative signals.
- **Missing elements in prior work:**
  a. No incremental sentence-by-sentence understanding.
  b. No integration of unstructured language with market state.
  c. No real-time, discourse-driven trading agents.

Data

To train and evaluate our model, we constructed a synchronized dataset of earnings call transcripts and high-frequency stock price data at 1-second intervals.

Data Sources:

- Data was collected from 4 earnings calls of each of 20 SP500 tech stocks
- Market data at 1-second resolution obtained via **Databento**
- Earnings call audio files sourced from **Capital IQ**

Preprocessing

- Converted audio to sentence level transcripts using **OpenAI Whisper**, with a **timestamp at the start of each sentence.**
- Aligned transcripts with **1-second stock price data** to create synchronized training data.
- Extrapolated second-by-second data to missing data points. This is typically a very safe assumption.

Training Data:

- For each sentence, the RL agent receives:
  - Current sentence embedding.
  - Recent stock price features.
  - Prior actions and rolling window of discourse.
- Agent trained to take **Buy / Hold / Sell** actions based on evolving discourse and market state.

This structured dataset enables our model to learn **dynamic trading policies** from **evolving financial discourse**, rather than relying on **predefined sentiment labels** or **static price-based strategies**."

ts_event	rtype	publisher_id	instrument_id	open	high	low	close	volume	symbol	is_extrapolated
2024-10-17 20:50:57	32	2	226	497.05	497.05	497.05	497.05	5	ADBE	FALSE
2024-10-17 20:50:58	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:50:59	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:00	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:01	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:02	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:03	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:04	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:05	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE
2024-10-17 20:51:06	32	2	226	497.05	497.05	497.05	497.05	0	ADBE	TRUE

Figure 1: Example stock data that is being fed to the model

```
Time (seconds)  Text
0.0     Good day and welcome to the Q3 FY24 Adobe earnings conference call.
6.0     Today's conference is being recorded.
9.0     At this time, I'd like to turn the conference over to Jonathan Voss, VP of Investor Relations.
14.0    Please go ahead.
15.0    Good afternoon and thank you for joining us.
19.0    With me on the call today are Shantanu Narayan, Adobe's Chair and CEO, David Woodwani, President
24.0    of Digital Media, and Mel Chakravarti, President of Digital Experience, and Dan Dern, Executive
29.0    Vice President and CFO.
32.0    On this call, which is being recorded, we will discuss Adobe's third quarter fiscal year
36.0    2024 financial results.
39.0    You can find our press release, as well as PDFs of our prepared remarks and financial
```

Figure 2: Transcript from a converted audio file with timestamps indicating the beginning of the sentence

Approach

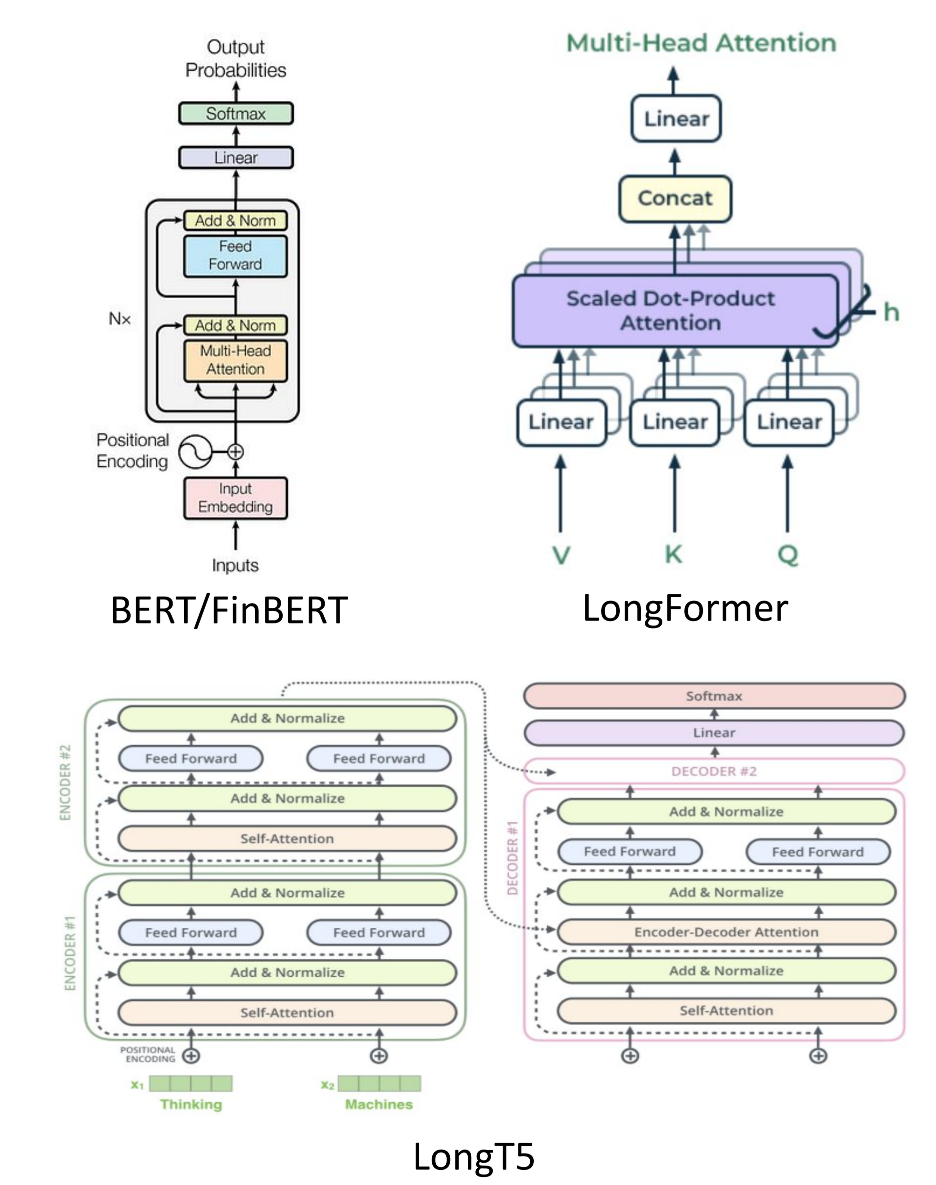

Figure 3: Transformer Architectures

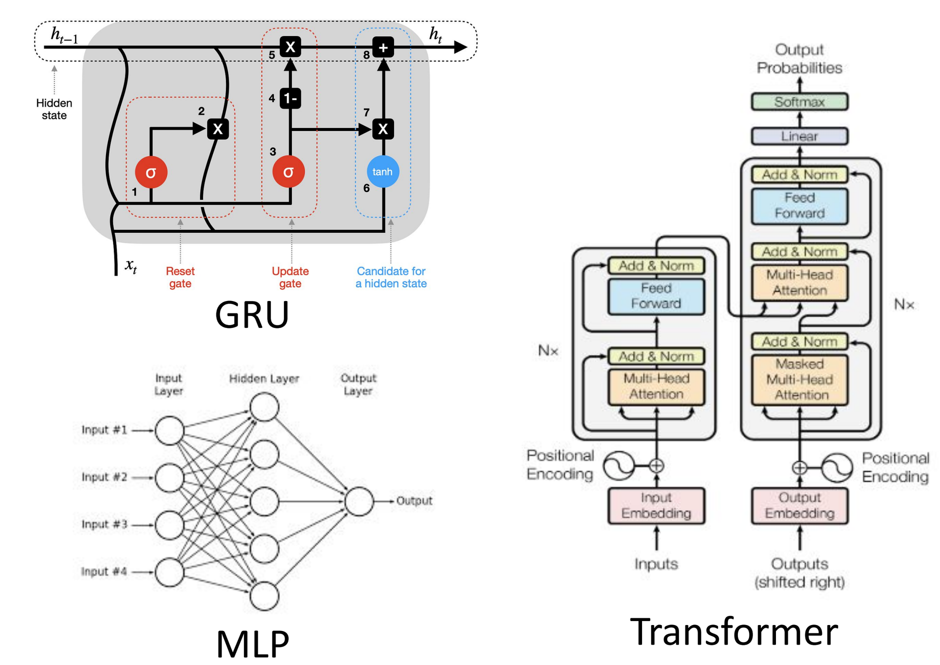

Figure 4: Policy Architectures

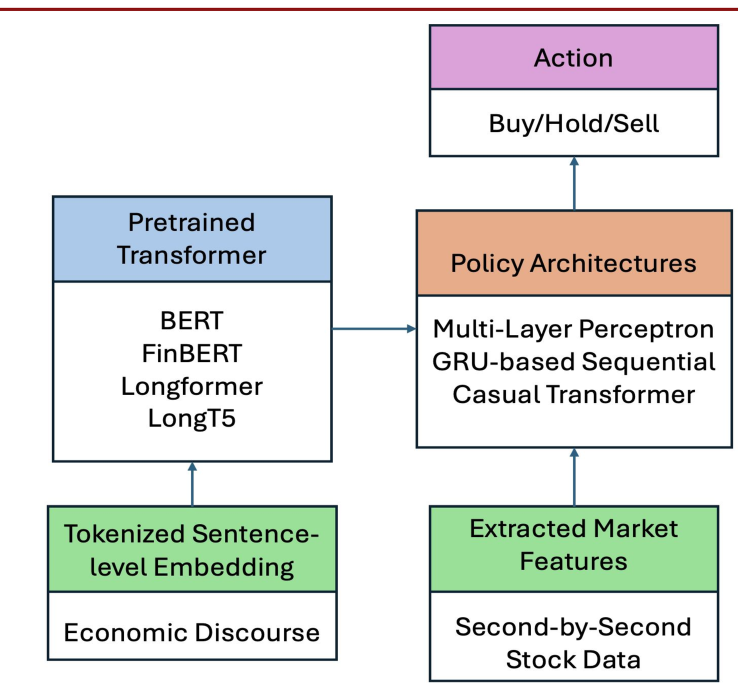

Figure 5: Model Architecture

Experiments

The goal of the experiment is to **evaluate** the effectiveness of **transformer-based reinforcement learning agents** in executing **real-time financial trades** based on **economic discourse**. Specifically, given a stream of financial text (e.g., earnings calls), the agent must map each sentence to a **buy, hold, or sell action**, using both **language and second-by-second market features**.

Inputs and Outputs:

- **Inputs:**
  - Sentence-level embeddings from **pretrained transformers** (BERT, FinBERT, Longformer, LongT5).
  - Synchronized **second-by-second stock price data** to reflect real market conditions.
- **Outputs:**
  - A sequence of **Buy / Hold / Sell actions**, one per sentence.
  - Performance metrics including **average return, Sharpe ratio, max/min return, and average drawdown**.

The RL models were trained using the Proximal Policy Optimization (PPO) algorithm for **10,000 timesteps**. The training utilized a **batch size of 64** and ran for **10 epochs per policy update**. The **Adam optimizer** was used with a **learning rate of 3×10^{-4}**. The policy and value networks were implemented as **two-layer multilayer perceptrons (MLPs)**, each with **64 hidden units per layer**, and **ReLU activations**. The agent interacted with a custom trading environment, where observations combined transformer-based text embeddings (from BERT, FinBERT, Longformer, or LongT5) and market features. The PPO agent used a **discount factor (γ) of 0.99** and a **GAE lambda of 0.95** for advantage estimation.

Model	Avg Return (%)	Avg Drawdown (%)	Avg Sharpe	Max Return (%)	Min Return (%)
BERT	-0.9	-3.25	-0.23	3.96	-5.29
FinBERT	-0.01	-0.54	-0.2	1.65	-0.48
Longformer	-0.75	-3.02	-0.23	4.62	-4.28
LongT5	-0.89	-3.29	-0.22	4.6	-5.34
FinBERT Static (Baseline)	0	-0.61	0.58	0.11	-0.5

Figure 6: Performance 4 Transformer Based RL Architures compared to FinBERT Static Baseline

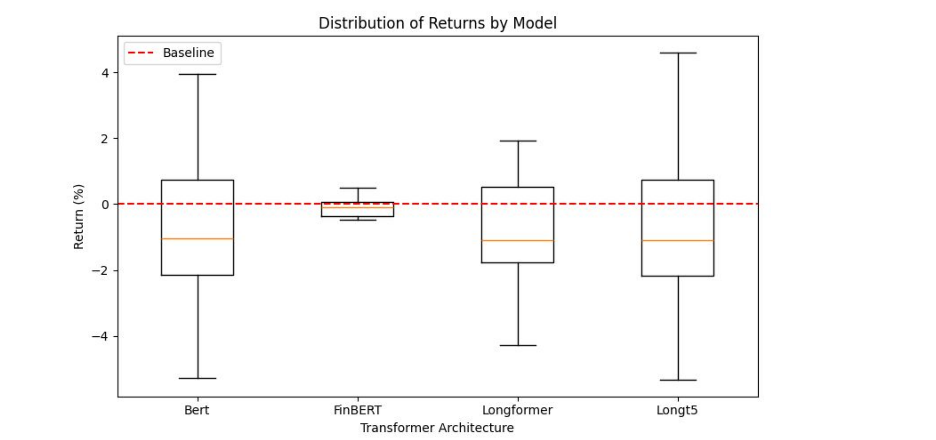

Figure 7: Distribution of returns from each Transformer Architecture

Analysis

FinBERT shows the most stable performance

- It achieved the lowest drawdown (-0.54%) and smallest min return (-0.48%), indicating more conservative behavior.

Transformer-based RL underperforms static FinBERT

- The static FinBERT classifier baseline outperforms all RL agents in return (0%) and Sharpe ratio (0.58).
- Likely due to MLP Policy architecture, which fails to capture temporal Nuance

Volatility is high in BERT, Longformer, and LongT5

- These models show wide return distributions (as seen in the boxplot) and large negative drawdowns (up to -5.34%).

Best upside: Longformer and LongT5

- Despite poor averages, these had the highest max returns (>4.6%), indicating potential if training is stabilized.
- Useful for higher-risk, event-driven strategies, but not yet consistent.

Conclusion

- **Transformer-based RL underperforms the static FinBERT baseline**, which had the highest Sharpe ratio (0.58) and no average drawdown, despite being the simplest model.
- **This underperformance is likely due to the MLP policy architecture**, which lacks temporal modeling and fails to capture evolving discourse dynamics.
- **Longformer and LongT5 offer high upside** as indicated by the volatility, suggesting they may be viable for event-driven or higher-risk trading strategies if stability can be improved.
- **Next steps: Improve training stability and implement GRU/Transformer based policy architecture** to capture temporal nuance. This would reduce the volatility displayed by Longformer and LongT5, and improve performance in comparison to the static baseline for all 4 transformer architectures..

References: Costola, Michele, et al. "Machine Learning Sentiment Analysis, COVID-19 News and Stock Market Reactions." *Research in International Business and Finance*, vol. 64, 1 Jan. 2023, p. 101881, www.sciencedirect.com/science/article/pii/S0275531923000077, https://doi.org/10.1016/j.ribaf.2023.101881. Accessed 22 Feb. 2023.