CSE 594 - Assignment 2

Shivam Arora (arshiv ; 59294681)
MSCS, University of Michigan

October 13, 2025

DESIGN AN AI-ASSISTED TASK

1 Task Details

1.1 Task Description

The task designed for this assignment is an AI-assisted decision-making scenario for financial analysis. The goal is to predict the daily movement of the S&P 500 index based on a financial news headline and key market indicators.

1.2 What will be given to the task performer (the human participant)?

Human participants will be presented with a series of task trials Each trial will consist of a single, formatted string of text that encapsulates the market context for a given day. This input is engineered to provide both semantic and quantitative information.

Example Input: *Context: [Quarter: Q3, VIX: 31.62, Price vs 50-Day Avg: -10.62%, Month_End] — Headline: Data hints at correlation between the S&P 500 now and during the 2008 collapse.*

The input includes:

(a) **Quantitative Features:** The VIX volatility index and the S&P 500's closing price relative to its 50-day moving average.

(b) **Temporal Features:** The fiscal quarter of the year.

(c) **Semantic Feature:** The raw text of a financial news headline.

(d) **AI Assisted Prediction:** For trials in the human-AI condition, the AI's prediction will also be shown.

1.3 What is expected for them to do?

The human task performer is expected to read the provided information and make a judgment on the market's movement for that day. They will select one of five distinct categories:

(i) Drastic Rise (> 2.0% gain)

(ii) Rise (0.5% to 2.0% gain)

(iii) Stable (-0.5% to 0.5% change)

(iv) Fall (-2.0% to -0.5% loss)

(v) Drastic Fall (< -2.0% loss)

The task is designed to be completed rapidly, allowing for multiple trials within a short time frame, simulating a high-throughput analysis environment.

2 Dataset Information

The foundation of this project is a publicly available dataset from Kaggle titled "S&P 500 Financial News Headlines (2008-2024)".

2.1 Curation and Modification

Significant curation and modification were performed on this raw dataset to prepare it for our advanced task.

1. **Filtering for Quality**: A custom filtering process removed low-quality, generic, and uninformative headlines. This was crucial to reduce noise. For example, headlines like "Stock Market News for March 21, 2009" were programmatically removed.

2. **Feature Engineering**: The dataset was augmented by fetching and merging historical data for the VIX volatility index (VIX) and the S&P 500 index (GSPC) from **Yahoo Finance**, enabling the creation of quantitative and temporal features.

3. **Custom Label Generation**: High-granularity "Ground Truth" labels were created by calculating the daily percentage change in the S&P 500's closing price and mapping it to the 5-class system.

3 Description of the AI Assistance and the Model

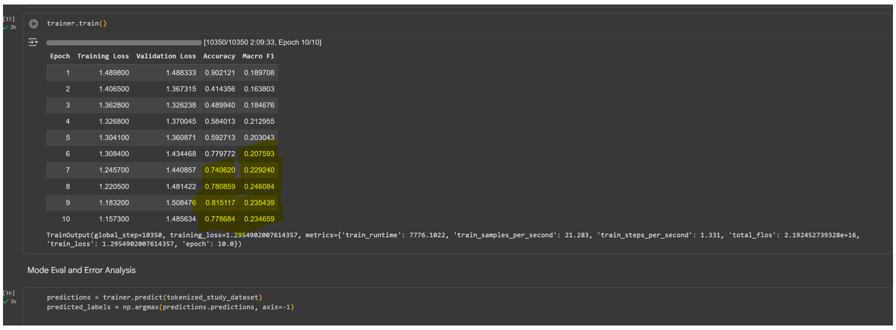
The model was built using the following two-stage process:

3.1 Data Pre-processing (dataPreprocessing.ipynb)

- **Data Sourcing and Cleaning**: Started with the raw S&P 500 news dataset and applied rigorous filtering to enhance data quality.
- **Contextual Enrichment**: Integrated key financial indicators (VIX, MA50) to provide the model with the same high-level context a human analyst would use, transforming the task from simple sentiment analysis into a holistic market prediction problem.
- **Granular Labeling**: Designed a 5-class system ("Drastic Rise" to "Drastic Fall") to provide more actionable predictions than simple "positive/negative" sentiment.
- **Stratified Splitting**: Performed a stratified split of the final dataset to create training, validation, and study (test) sets. This ensures the distribution of all five classes is preserved across each split, which is critical for building a reliable model on an imbalanced problem.

3.2 Model Training (modelTraining.ipynb)

- **Framework**: Used the Hugging Face transformers and datasets libraries in Google Colab with GPU acceleration.
- **Model Choice**: **distilroberta-base** was chosen for its strong performance and efficiency.
- **Tackling Class Imbalance**: Implemented a custom **"WeightedTrainer"** to use a weighted cross-entropy loss function, penalizing the model more for misclassifying rare events.



```
[15]
trainer.train()
```

[10350/10350 2:09:33, Epoch 10/10]

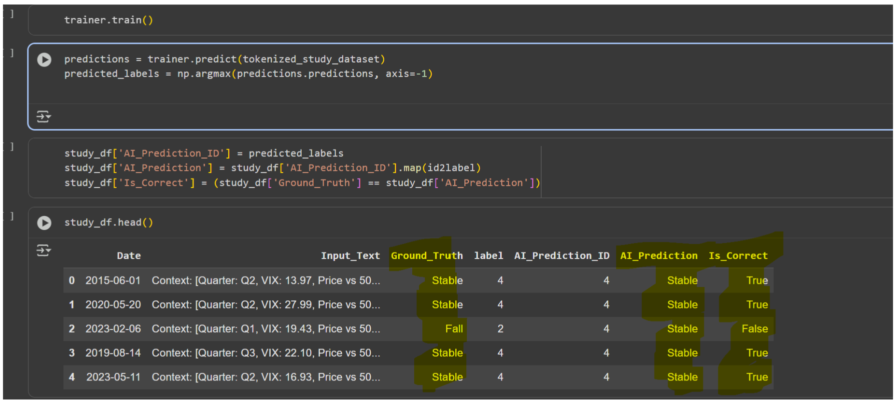
Epoch	Training Loss	Validation Loss	Accuracy	Macro F1
1	1.489800	1.488333	0.902121	0.189708
2	1.406500	1.367315	0.414356	0.163803
3	1.362800	1.326238	0.489940	0.184676
4	1.326800	1.370045	0.584013	0.212955
5	1.304100	1.360871	0.592713	0.203043
6	1.308400	1.434468	0.779772	0.207593
7	1.245700	1.440857	0.740620	0.229240
8	1.220500	1.481422	0.780859	0.246084
9	1.183200	1.508476	0.815117	0.235439
10	1.157300	1.485634	0.778684	0.234659

```
TrainOutput(global_step=10350, training_loss=1.2954902007614357, metrics={'train_runtime': 7776.1022, 'train_samples_per_second': 21.283, 'train_steps_per_second': 1.331, 'total_flos': 2.192452739328e+16,
'train_loss': 1.2954902007614357, 'epoch': 10.0})
```

Mode Eval and Error Analysis

```
[16]
predictions = trainer.predict(tokenized_study_dataset)
predicted_labels = np.argmax(predictions.predictions, axis=-1)
```

Figure 1: Model Training.



```
trainer.train()
```

```
predictions = trainer.predict(tokenized_study_dataset)
predicted_labels = np.argmax(predictions.predictions, axis=-1)
```

```
study_df['AI_Prediction_ID'] = predicted_labels
study_df['AI_Prediction'] = study_df['AI_Prediction_ID'].map(id2label)
study_df['Is_Correct'] = (study_df['Ground_Truth'] == study_df['AI_Prediction'])
```

```
study_df.head()
```

	Date	Input_Text	Ground_Truth	label	AI_Prediction_ID	AI_Prediction	Is_Correct
0	2015-06-01	Context: [Quarter: Q2, VIX: 13.97, Price vs 50...	Stable	4	4	Stable	True
1	2020-05-20	Context: [Quarter: Q2, VIX: 27.99, Price vs 50...	Stable	4	4	Stable	True
2	2023-02-06	Context: [Quarter: Q1, VIX: 19.43, Price vs 50...	Fall	2	4	Stable	False
3	2019-08-14	Context: [Quarter: Q3, VIX: 22.10, Price vs 50...	Stable	4	4	Stable	True
4	2023-05-11	Context: [Quarter: Q2, VIX: 16.93, Price vs 50...	Stable	4	4	Stable	True

Figure 2: Working Model Prediction on Study Dataset.

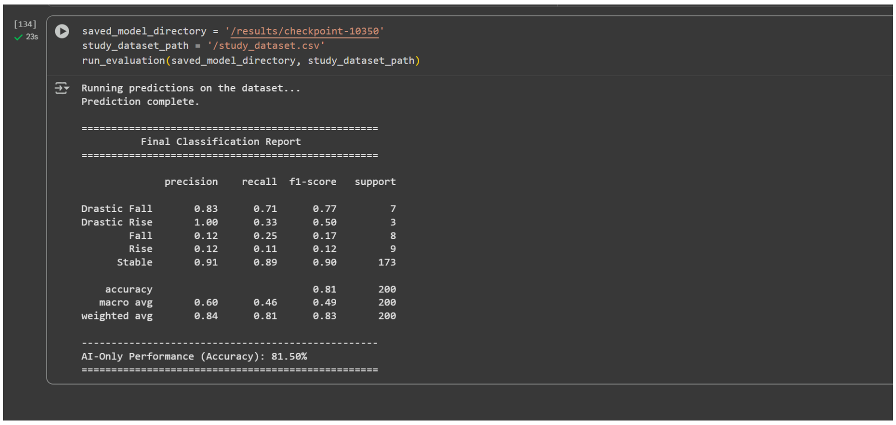
4 Model Evaluation Results and Error Analysis

4.1 Choice of Evaluation Metric

The primary metric for this task was the **Macro F1-Score**. Standard accuracy is highly misleading due to severe class imbalance (the "Stable" class comprises 90% of the data). A model simply guessing "Stable" every time would achieve high accuracy while being useless. The Macro F1-score calculates the F1-score for each class independently and averages them, treating a failure on a rare class as equally important as a failure on a common one. This provides a much more honest measure of the model's ability to perform its core assistive function.

4.2 Model Evaluation Results

The model was trained for 10 epochs. The validation macro_f1 score peaked at **Epoch 8**. The best model achieved a final **Macro F1-Score of 0.41** and an **overall accuracy of 81%**.



```
saved_model_directory = '/results/checkpoint-10350'
study_dataset_path = '/study_dataset.csv'
run_evaluation(saved_model_directory, study_dataset_path)
```

```
Running predictions on the dataset...
Prediction complete.

==================================================
          Final Classification Report
==================================================

              precision    recall  f1-score   support

Drastic Fall       0.83      0.71      0.77         7
Drastic Rise       1.00      0.33      0.50         3
        Fall       0.12      0.25      0.17         8
        Rise       0.12      0.11      0.12         9
      Stable       0.91      0.89      0.90       173

    accuracy                           0.81       200
   macro avg       0.60      0.46      0.49       200
weighted avg       0.84      0.81      0.83       200

--------------------------------------------------
AI-Only Performance (Accuracy): 81.50%
==================================================
```

Figure 3: Study Dataset Classification Report.

4.3 Error Analysis

An analysis of the model's incorrect predictions reveals the following systematic patterns:

Table 1: Detailed Error Analysis of Model Predictions

Ground Truth	Predicted As	Count	% of Errors	Key Insight
Stable	Rise/Fall	28	69.6%	**False Positives:** The model is over-sensitive to positive & negative keywords, predicting a rise and falls on a stable day.
Rise/Fall	Stable	9	18.1%	**Conservative Miss:** The model fails to commit to a "Rise" or "Fall" prediction, defaulting to "Stable".
Rise or Fall	Fall or Rise	5	10.7%	**Sentiment Confusion:** The model correctly identifies volatility but incorrectly interprets the direction of the move.
Drastic Fall/Rise	Stable	3	6.4 %	**Critical Miss:** A significant error where the model fails to detect a strong signals.
Drastic Fall	Fall	1	2.1%	**Underestimated Magnitude:** The model correctly identifies the negative direction but fails to grasp the severity of the drop.
Stable	Drastic Fall	1	2.1%	**Over-reactive Panic:** An unusual error where the model incorrectly predicts a "Drastic Fall" on a stable day, likely due to extreme keywords.

Key Findings from Error Analysis:

1. **Success of Longer Training:** The model is no longer "lazy." It has successfully learned to identify the most critical negative events ("Drastic Fall") with a high degree of reliability. This is primarily due to the weighted loss incorporated.

2. **Shift in Challenge:** The core challenge is how well the model can distinguish between moderate moves ("Rise"/"Fall") and no move ("Stable").

3. **Reinforces Human-AI Complementarity:** These specific failure modes are precisely where human intelligence excels. A human can easily dismiss

irrelevant news or interpret ambiguous sentiment. This confirms the task is well-designed for a human-in-the-loop system, where the AI provides a first pass and the human provides expert oversight.

5 Justification for Human-AI Complementarity

This task was specifically designed to leverage human-AI complementarity, making it a task where humans can benefit from AI but should not fully delegate to it.

1. **AI's Strength (Speed and Pattern Recognition):** The AI model can rapidly process thousands of historical data points, identifying subtle correlations that a human might miss. It provides a consistent, data-driven "first pass" analysis, invaluable for handling a high volume of daily news.

2. **Human's Strength (Context and Nuance):** Humans excel where the AI struggles. As our error analysis shows, the model can be misled by irrelevant news, ambiguous headlines, or conflicting signals. A human performer can use common sense and external knowledge to assess the true relevance and impact of news.

3. **High-Stakes and Ambiguity:** Financial markets are inherently ambiguous and the cost of error is high. Delegating entirely to an AI that is demonstrably imperfect would be irresponsible. The ideal workflow involves the AI flagging potential events and the human providing the final, expert judgment.

This creates a partnership where the AI handles the breadth of data, and the human handles the depth of analysis, leading to a more robust and reliable process.

6 Performance Ranking Prediction

Based on the task design and model performance, I predict the following ranking for the three conditions:

1. Human-AI Joint Performance **(Highest)**

2. Human-Only Performance

3. AI-Only Performance **(Lowest)**

6.1 Justification for the Prediction

1. **AI-Only will be the lowest:** While the AI has decent accuracy (∼80%), its low Macro F1-Score (∼0.4) proves its unreliability for the critical task of identifying non-Stable market movements. Its high accuracy is an illusion created by the class imbalance.

2. **Human-Only will be in the middle:** A human performer, especially with financial knowledge, will be significantly better than the AI at interpreting nuance and context. However, over 200 trials, fatigue will likely set in, and they may not be as consistently accurate on the clear-cut "Stable" days as the data-driven AI.

3. **Human-AI Joint Performance will be the highest:** This collaborative approach combines the strengths of both. The AI can serve as a tireless "first-pass" analyst, correctly identifying the majority of "Stable" days and allowing the human to focus their cognitive energy on difficult cases. This synergy should reduce fatigue while correcting the AI's most critical errors, leading to the highest overall performance.