Hybrid Sentiment-Price Prediction: Leveraging FinBERT and LSTM for Stock Price Movement Forecasting

Presenter: Reza Tashakkori

UNIVERSITY OF LUXEMBOURG

Objectives and Methodology

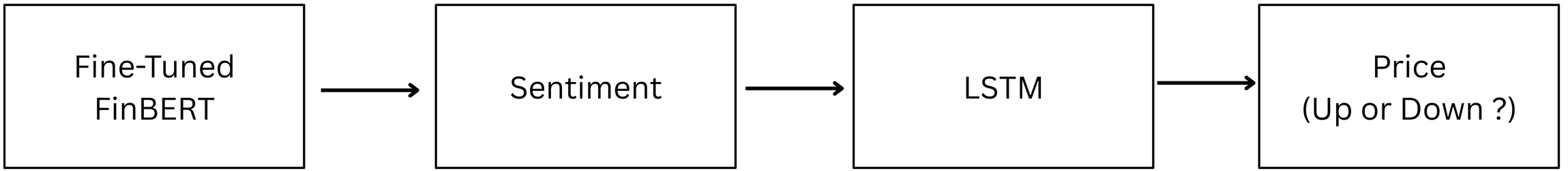
Aim: Predict stock price changes (Up/Down) for 4 companies (Tesla, Apple, Microsoft, NVIDIA) using financial news sentiment and stock price data from **2020–2024**.

Approach:

1. **Fine-Tuned FinBERT** on financial news to classify sentiment (0: Negative, 1: Neutral, 2: Positive)
2. Train an **LSTM** model on price sequences and sentiment features for binary price prediction (Up/Down)
3. Combine FinBERT and LSTM into a hybrid model: using FinBERT to generate sentiment labels, and integrating these with price sequences as input to a pre-trained LSTM, achieving 82.47% accuracy in a unified prediction process.

Recap: Merges textual sentiment data and numerical price data into one learning framework.

Hybrid Model



Datasets Overview

FNSPID Dataset (NASDAQ Financial News, Post-2015):

- Features: Date (y-m-d), News Title, Stock Symbol.
- Applied FinBERT to generate Sentiment Labels (0: Negative, 1: Neutral, 2: Positive).
- 5 Samples of this Dataset:

	Article_title	Date	Stock_symbol	Sentiment_Label	Company
31841	Tesla supplier Syrah expects more graphite buy...	2023-10-25	TSLA	1	Tesla
30039	Akamai (AKAM) to Report Q4 Earnings: What's in...	2022-02-10	NVDA	1	NVIDIA
2393	Chipmaker Skyworks sees profit below estimates...	2023-08-07	AAPL	0	Apple
8955	Got $1,000? 5 Buffett Stocks to Buy and Hold F...	2023-12-13	AMZN	2	Amazon
21495	Dow Analyst Moves: MSFT	2022-07-18	MSFT	1	Microsoft

yFinance API Dataset (Stock Data, Post-2015):

- Features: Date (y-m-d), Close Price, Price Change (0: Down, 1: Up), Stock Symbol.
- 5 Samples of this Dataset:

	Close_Price	Price_Change	Stock_symbol	Date	Company
13990	51.730362	1	MSFT	2016-08-12	Microsoft
4770	15.559333	0	TSLA	2015-12-17	Tesla
14317	77.060814	1	MSFT	2017-11-29	Microsoft
12115	41.237312	1	AAPL	2018-03-01	Apple
15528	235.949722	0	MSFT	2022-09-22	Microsoft

Merging Process:

- Aggregated FNSPID by Date and Stock Symbol, calculating Avg_Sentiment (mean of Sentiment Labels).
- Merged with yFinance on Date and Stock Symbol (inner join).
- Added original news titles and Sentiment Labels back into the merged dataset.

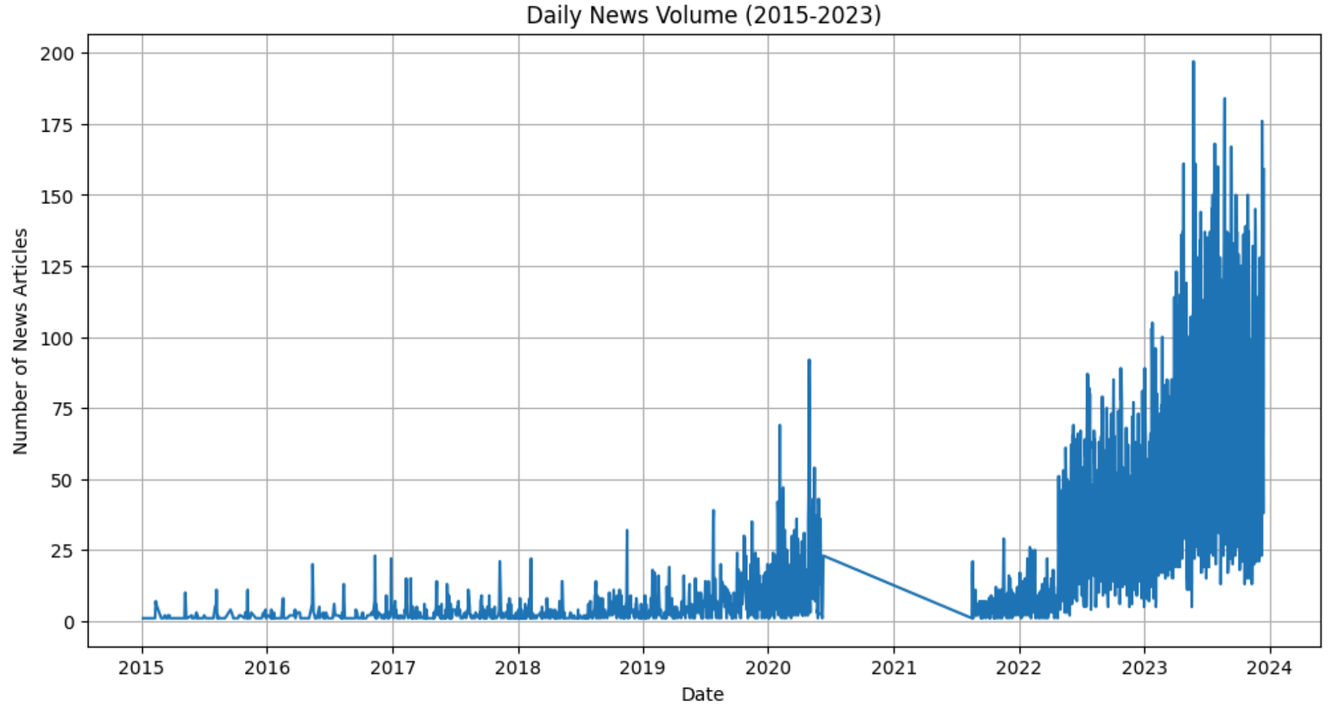
Initial Merged Dataset (6 Companies: TSLA, AAPL, MSFT, NVDA, GOOGL, AMZN; 2015-01 to 2023-12)

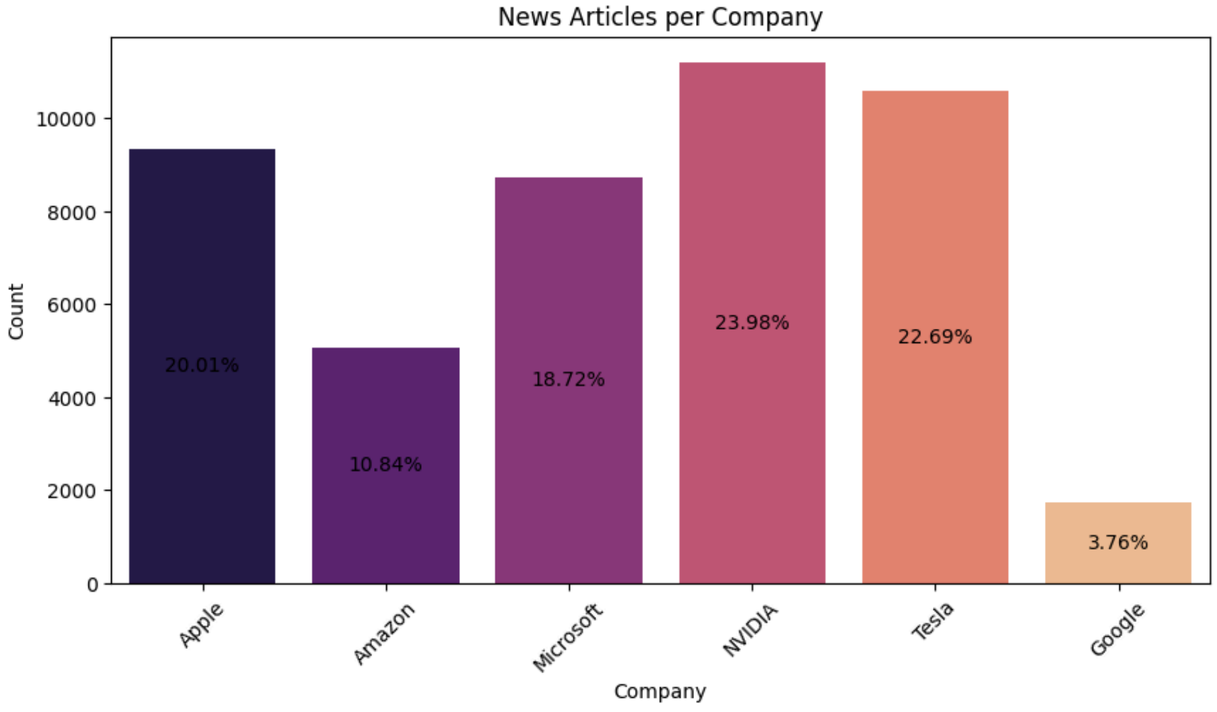
- Shape: 42589 rows, 8 columns

Final Merged Dataset (4 Companies: TSLA, AAPL, MSFT, NVDA; 2020-01 to 2023-12):

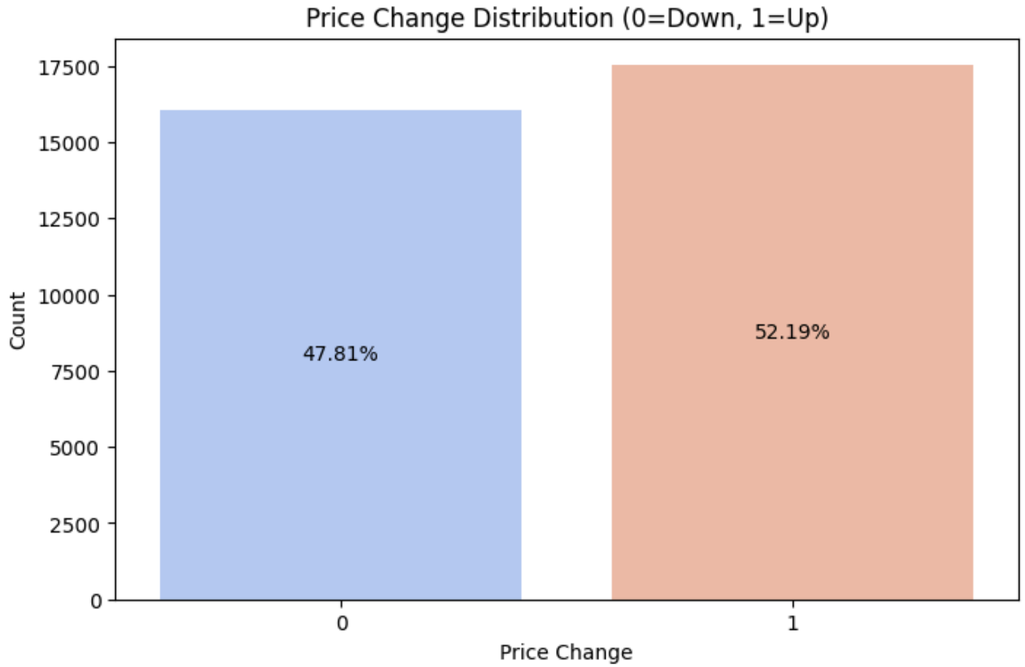
- Removed Google (GOOGL) and Amazon (AMZN) due to class imbalance.
- Shortened the date range to reduce FinBERT fine-tuning runtime (from ~12 hours to 1 hour).
- Shape: 33,595 rows, 7 columns.
- No missing values!
- 5 Sample of final Dataset:

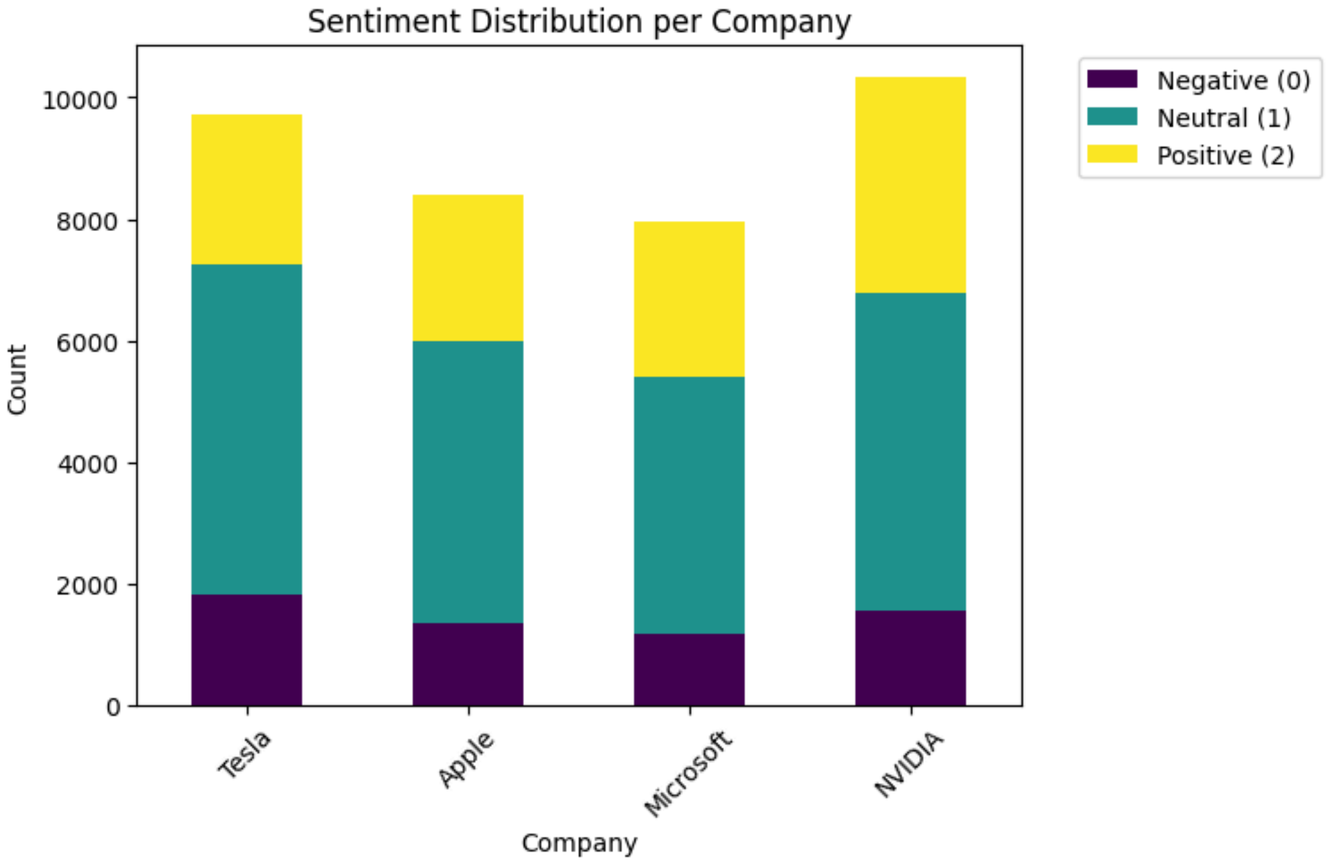
	Date	Company	Article_title	Sentiment_Label	Close_Price	Price_Change	Avg_Sentiment
23961	2023-09-22	Microsoft	Businesses Will Invest $200 Billion in AI by 2...	2	313.391296	1	1.185185
18892	2022-09-23	Microsoft	Wall Street's banking-as-a-service has a problem	0	232.953598	0	0.642857
11990	2022-12-06	Apple	Is Snowflake Stock A Buy Following Q3 Results?	2	141.082489	0	1.000000
22350	2023-06-15	Microsoft	AI, ESG Come Together in This ETF	1	343.399384	0	1.066667
28333	2022-11-30	NVIDIA	US STOCKS-Wall Street rises after Powell eyes ...	1	16.909143	1	1.125000

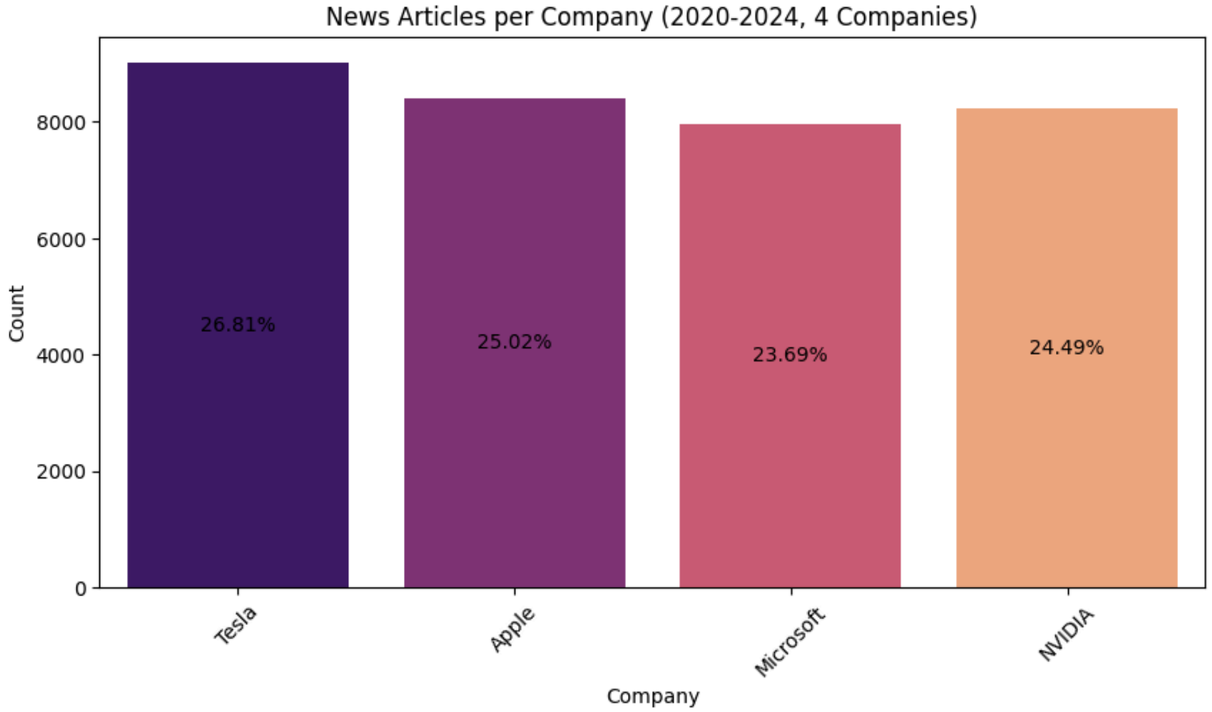




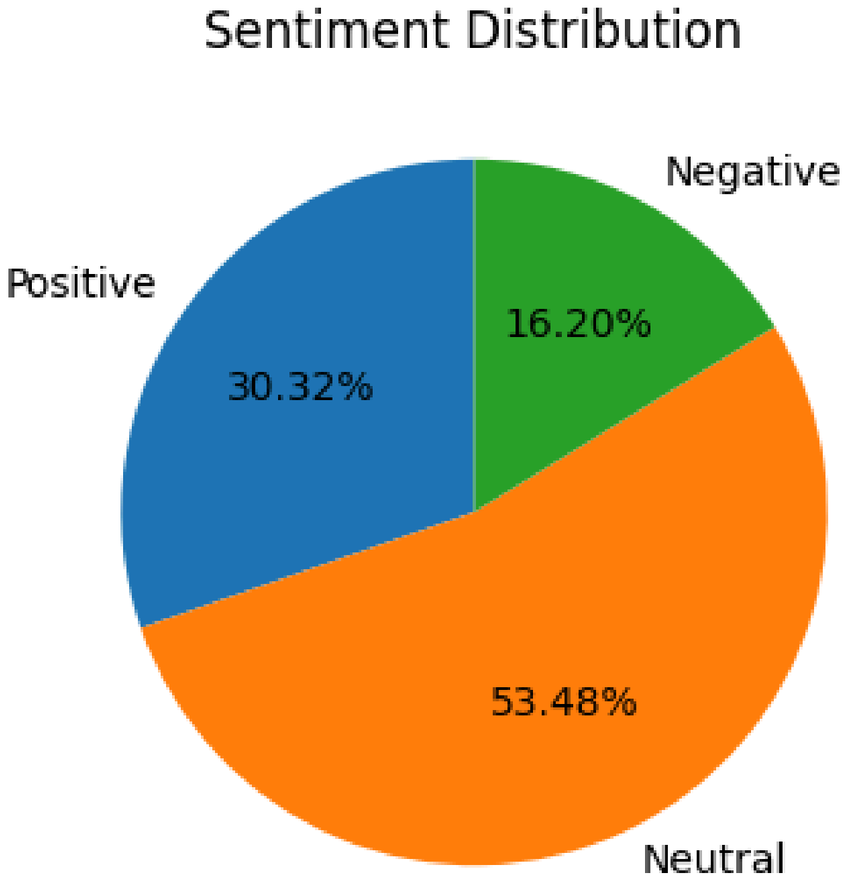
Exploring the Datasets

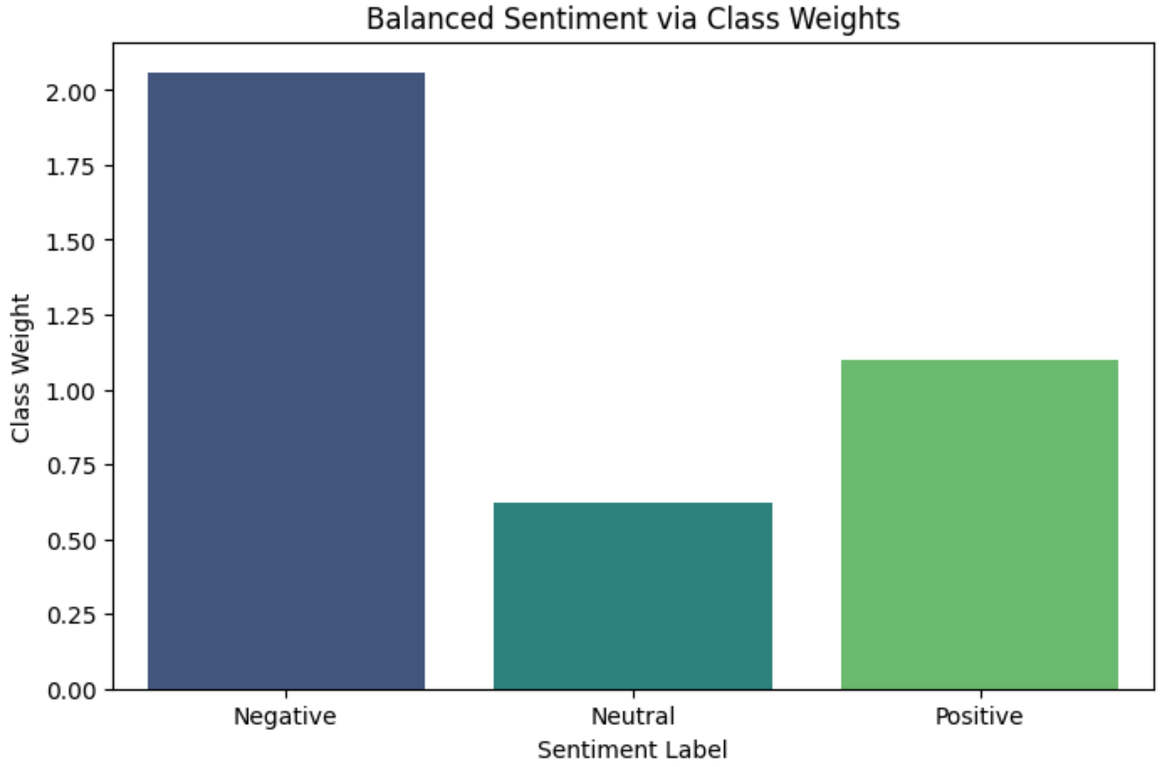






More ...


Sentiment Distribution
Negative
Positive
16.20%
30.32%
53.48%
Neutral


Balanced Sentiment via Class Weights
Class Weight
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Negative
Neutral
Positive
Sentiment Label

Fine-Tuning FinBERT for Sentiment Analysis

Model Overview and Fine-Tuning:

- **BERT** is a language model developed by Google that reads text in both directions to understand context.
- **FinBERT**, is BERT-based language model pre-trained on financial texts to perform sentiment analysis in the finance domain.
- It uses a 12-layer Transformer with 768 hidden units, 12 attention heads, and 110M parameters to classify in 3 classes (0: Negative, 1: Neutral, 2: Positive).
- **Fine-Tuned FinBERT** has optimized the weights for our final dataset and specific task, retaining the same structure for optimized performance.

Key Training Components:

- **Dataset Split**: Stratified by company into train (80%), val (10%), and test (10%) for balanced representation across companies.
- **Loss Function**: Applied weighted cross-entropy to address class imbalance, improving fairness in predictions.
- **Training Settings**:
  - Trained for 5 epochs (sufficient for convergence in this case)
  - Batch Size 32 (balanced efficiency and memory)
  - Learning Rate 2e-5 (standard for fine-tuning)
  - 500 Warmup Steps (slowly increase lr for stable learning)
  - 0.01 Weight Decay (penalty for large weigths to prevents overfitting)
  - Early Stopping (Patience = 3) (avoids overtraining)

Evaluate Fine-Tuned FinBERT Performance

Results Summary:

- **Test Accuracy**: 95.24%, with per-class accuracies (Negative: 93.98%, Neutral: 95.01%, Positive: 96.34%).
- **Runtime**: 64.47 minutes on T4-GPU, model saved for future use.

Analysis:

- Accuracy of 95.24% and F1 scores (0.94–0.96) across classes show robust performance.
- Val accuracy grew from 92.38% to 95.15% over 5 epochs, indicating sufficient convergence.
- Val loss increased after epoch 2 (0.15653 to 0.29453), suggesting potential mild overfitting.

Concern:

- Training for 5 epochs may overfit, as val loss rose after epoch 2.
- Stopping earlier (e.g., after epoch 3) could reduce overfitting risks - more generalizable.
- High runtime (64.47 minutes) reflects computational intensity, manageable with GPU optimization.

```
Test Accuracy: 95.24%
Class Negative Test Accuracy: 93.98%
Class Neutral Test Accuracy: 95.01%
Class Positive Test Accuracy: 96.34%
Confusion Matrix (rows: true, cols: predicted):
          Negative  Neutral  Positive
Negative       531       20        14
Neutral         29     1695        60
Positive         9       28       974
Classification Report:
              precision    recall  f1-score   support

    Negative       0.93      0.94      0.94       565
     Neutral       0.97      0.95      0.96      1784
    Positive       0.93      0.96      0.95      1011

    accuracy                           0.95      3360
   macro avg       0.95      0.95      0.95      3360
weighted avg       0.95      0.95      0.95      3360

Total Runtime: 64.47 minutes (using T4-GPU Google Colab)
Model saved as finbert_finetuned.pth
```

```
Epoch 1/5, Train Loss: 0.14405, Val Loss: 0.19376, Train Acc: 94.52299%, Val Acc: 92.37868%, Val Precision: 0.92804, Val Recall: 0.92379, Val F1: 0.92433
Best model saved at epoch 1
Epoch 2/5, Train Loss: 0.04845, Val Loss: 0.15653, Train Acc: 98.45587%, Val Acc: 93.89699%, Val Precision: 0.93943, Val Recall: 0.93897, Val F1: 0.93909
Best model saved at epoch 2
Epoch 3/5, Train Loss: 0.02831, Val Loss: 0.22863, Train Acc: 99.07724%, Val Acc: 93.80768%, Val Precision: 0.93962, Val Recall: 0.93808, Val F1: 0.93823
Best model saved at epoch 3
Epoch 4/5, Train Loss: 0.01404, Val Loss: 0.27515, Train Acc: 99.61676%, Val Acc: 94.52218%, Val Precision: 0.94601, Val Recall: 0.94522, Val F1: 0.94529
Best model saved at epoch 4
Epoch 5/5, Train Loss: 0.00666, Val Loss: 0.29453, Train Acc: 99.81024%, Val Acc: 95.14737%, Val Precision: 0.95168, Val Recall: 0.95147, Val F1: 0.95152
Best model saved at epoch 5
```

LSTM Model for Price Prediction

Model Selection and Overview:

- **LSTM** (as an improved RNN for remembering longer-term patterns) for its ability to capture sequential patterns in time-series data, ideal for stock price prediction.
- Compared to GRU (faster and simpler) and CNNs (both <75% accuracy), **LSTM** was best with **85.80%** accuracy.
- LSTM outperforms basic RNNs by preventing vanishing gradients, modeling sequences well, and suiting not large datasets with low compute needs (vs. Transformers)
- Designed a 3-layer LSTM with 512 hidden units, processing 12 input features over 3-day sequences for binary classification (Up/Down), with a 0.3 dropout rate to prevent overfitting (**many-to-one RNN**).

Key Training Components:

- **Data Preparation**: Preprocessed data with lagged and rolling features (e.g., 5-day averages) to capture trends, split into train (60%), validation (20%), and test (20%) sets with stratification by company for balanced representation.

- **Loss and Optimization**: Used weighted cross-entropy with class weights to handle imbalance, trained with Adam Optimizer (lr=0.003) (not RMSProp - may not generalize properly), and a Learning Rate Scheduler (prevents overshooting or getting stuck) to ensure stable convergence.

- **Training Settings**: Trained for up to 50 epochs with a Batch Size of 128, Early Stopping (Patience=5) on Validation Loss, and Weight Decay 0.01 to enhance generalization, ensuring robust performance.

Assess LSTM Performance

Results Summary:

- **Test Accuracy**: 85.80%, with per-class accuracies (Down: 83.17%, Up: 88.21%).
- **Runtime**: 1.19 minutes on T4-GPU, model saved for future use.

Analysis:

- Test Accuracy 85.80%, with consistent F1 scores (0.85–0.87), showing good prediction ability.
- Validation accuracy improved from 51.69% to 86.14%, and loss dropped from 0.69331 to 0.34575, with a small train-val gap (0.26 vs. 0.35), indicating solid convergence.

Concerns:

- Fast runtime (1.19 min) is efficient, but 50 epochs might not capture subtle patterns
- Early stopping didn't trigger, suggesting more epochs could enhance learning.
- **BUT!**, Validation loss stopped improving (stuck at 0.34575)—further training will be useless!

```
Test Accuracy: 85.80%
Class Down Test Accuracy: 83.17%
Class Up Test Accuracy: 88.21%
Confusion Matrix (rows: true, cols: predicted):
      Down    Up
Down  2669   540
Up     414  3096
Classification Report:
              precision    recall  f1-score   support

        Down       0.87      0.83      0.85      3209
          Up       0.85      0.88      0.87      3510

    accuracy                           0.86      6719
   macro avg       0.86      0.86      0.86      6719
weighted avg       0.86      0.86      0.86      6719

Total Runtime: 1.19 minutes (using T4-GPU Google Colab)
Model saved as price change lstm best.pth
```

```
Epoch 1/50, Train Loss: 0.68884, Val Loss: 0.69331, Train Acc: 52.43624%, Val Acc: 51.68924%, Val Precision: 0.52199, Val Recall: 0.81085, Val F1: 0.63512
Epoch 2/50, Train Loss: 0.68639, Val Loss: 0.68737, Train Acc: 52.42136%, Val Acc: 51.85295%, Val Precision: 0.51853, Val Recall: 1.00000, Val F1: 0.68294
Epoch 3/50, Train Loss: 0.68595, Val Loss: 0.68911, Train Acc: 52.94731%, Val Acc: 52.90966%, Val Precision: 0.54149, Val Recall: 0.59931, Val F1: 0.56894
Epoch 4/50, Train Loss: 0.68769, Val Loss: 0.69018, Train Acc: 52.96219%, Val Acc: 53.17756%, Val Precision: 0.52743, Val Recall: 0.93284, Val F1: 0.67385
Epoch 5/50, Train Loss: 0.68411, Val Loss: 0.68022, Train Acc: 53.92974%, Val Acc: 54.36821%, Val Precision: 0.53422, Val Recall: 0.93657, Val F1: 0.68036
```

...

```
Epoch 45/50, Train Loss: 0.27857, Val Loss: 0.34941, Train Acc: 87.77910%, Val Acc: 86.24795%, Val Precision: 0.85935, Val Recall: 0.87859, Val F1: 0.86886
Epoch 46/50, Train Loss: 0.28467, Val Loss: 0.35165, Train Acc: 87.34742%, Val Acc: 85.72704%, Val Precision: 0.84168, Val Recall: 0.89265, Val F1: 0.86642
Epoch 47/50, Train Loss: 0.26897, Val Loss: 0.34913, Train Acc: 88.31001%, Val Acc: 86.06936%, Val Precision: 0.85567, Val Recall: 0.87974, Val F1: 0.86753
Epoch 48/50, Train Loss: 0.26657, Val Loss: 0.34850, Train Acc: 88.46879%, Val Acc: 86.14377%, Val Precision: 0.86840, Val Recall: 0.86366, Val F1: 0.86602
Epoch 49/50, Train Loss: 0.26421, Val Loss: 0.35088, Train Acc: 88.45390%, Val Acc: 85.66751%, Val Precision: 0.83875, Val Recall: 0.89581, Val F1: 0.86634
Epoch 50/50, Train Loss: 0.26225, Val Loss: 0.34575, Train Acc: 88.59780%, Val Acc: 86.14377%, Val Precision: 0.84963, Val Recall: 0.89036, Val F1: 0.86952
```

Hybrid/Combined Model Integration

Model Overview:

- Combines Fine-Tuned FinBERT for sentiment labeling and a pre-trained LSTM for price sequences, achieving **82.47%** accuracy.
- Merges text and price data, using FinBERT's sentiment skills and LSTM's sequence handling.

Caution: Weights match the LSTM model; difference lies in using FinBERT-generated sentiment features.

Key Code Components:

- **Model Loading**: Loads pre-trained FinBERT and LSTM models, ensuring no retraining and leveraging prior optimization.
- **Data Preprocessing**: Generates sentiment labels and features, merging with price data for trend capture.

Hybrid Model Results and Justification

Results and Analysis:

- Achieved 82.47% Test Accuracy (Down: 78.80%, Up: 85.76%), with F1-Score (0.84) and ROC-AUC (0.90), showing good prediction.
- **Runtime** of 3.72 minutes on T4-GPU reflects efficiency.
- High ROC-AUC and balanced F1 indicate robust performance.
- **Concerns**: Slight accuracy drop (82.47% vs. LSTM's 85.80%) suggests sentiment integration mismatch;

Justification:

- Results outperform earlier models (GRU and CNNs < 75%) with good F1 scores (>80%), ROC-AUC (~90%), and a reliable confusion matrix, showcasing the hybrid approach's value.
- **MCC** (0.65) shows moderately strong performance with limitations in handling imbalanced classes, yet the model effectively handles financial prediction complexity.

```
Combined Model Test Accuracy: 82.47%
Precision: 0.82, Recall: 0.86, F1-Score: 0.84, ROC-AUC: 0.90, MCC: 0.65
Class Down Test Accuracy: 78.80%
Class Up Test Accuracy: 85.76%
Confusion Matrix (rows: true, cols: predicted):
      Down    Up
Down  2502   673
Up     504  3036
Classification Report:
              precision    recall  f1-score   support

        Down       0.83      0.79      0.81      3175
          Up       0.82      0.86      0.84      3540

    accuracy                           0.82      6715
   macro avg       0.83      0.82      0.82      6715
weighted avg       0.83      0.82      0.82      6715

Total Runtime: 3.72 minutes (using T4-GPU Google Colab)
Combined model saved as /content/combined_model.pth
```

* **MCC** (Matthews Correlation Coefficient) is a performance metric used to evaluate binary or multi-class classification models, especially when classes are imbalanced (+1 = perfect prediction, 0 = random guessing, -1 = inverse prediction).

Real-World Applications

Prediction Power:

- Forecasts stock price movements (Up/Down) at 82.47% accuracy.
- Leverages Fine-Tuned FinBERT (95.24% accurate) for real-time news sentiment.
- Uses LSTM (85.80% accurate) for historical price data.
- Supports traders and analysts with reliable insights.

Adaptability:

- Adapts to fresh news data effortlessly.
- Outperforms static sentiment models.

Scalability:

- Extends to new companies or markets.
- Works with updated news and price inputs.

Thank you :)