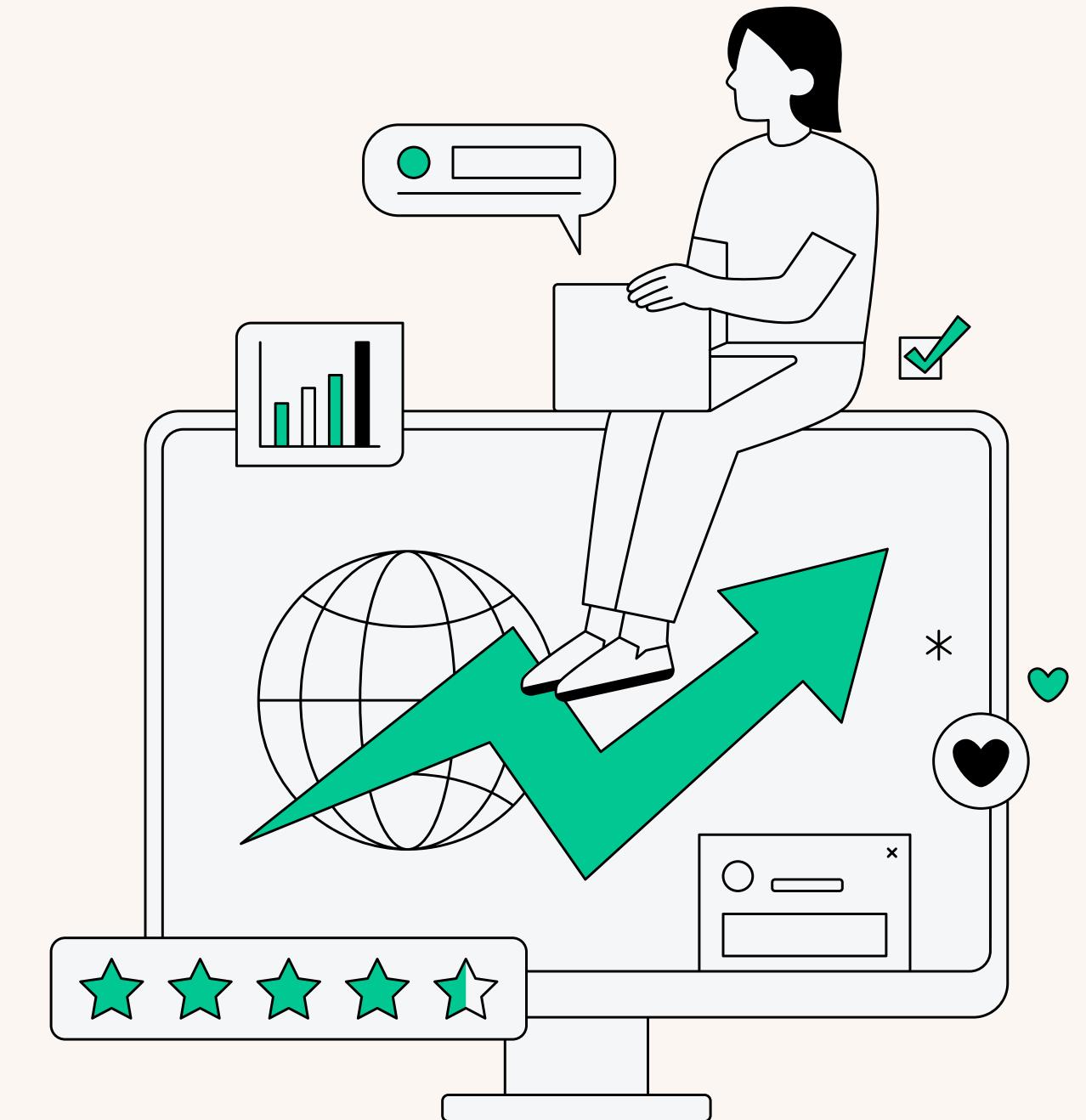


Presented by Stemtoklem

STOCK PREDICTIONS

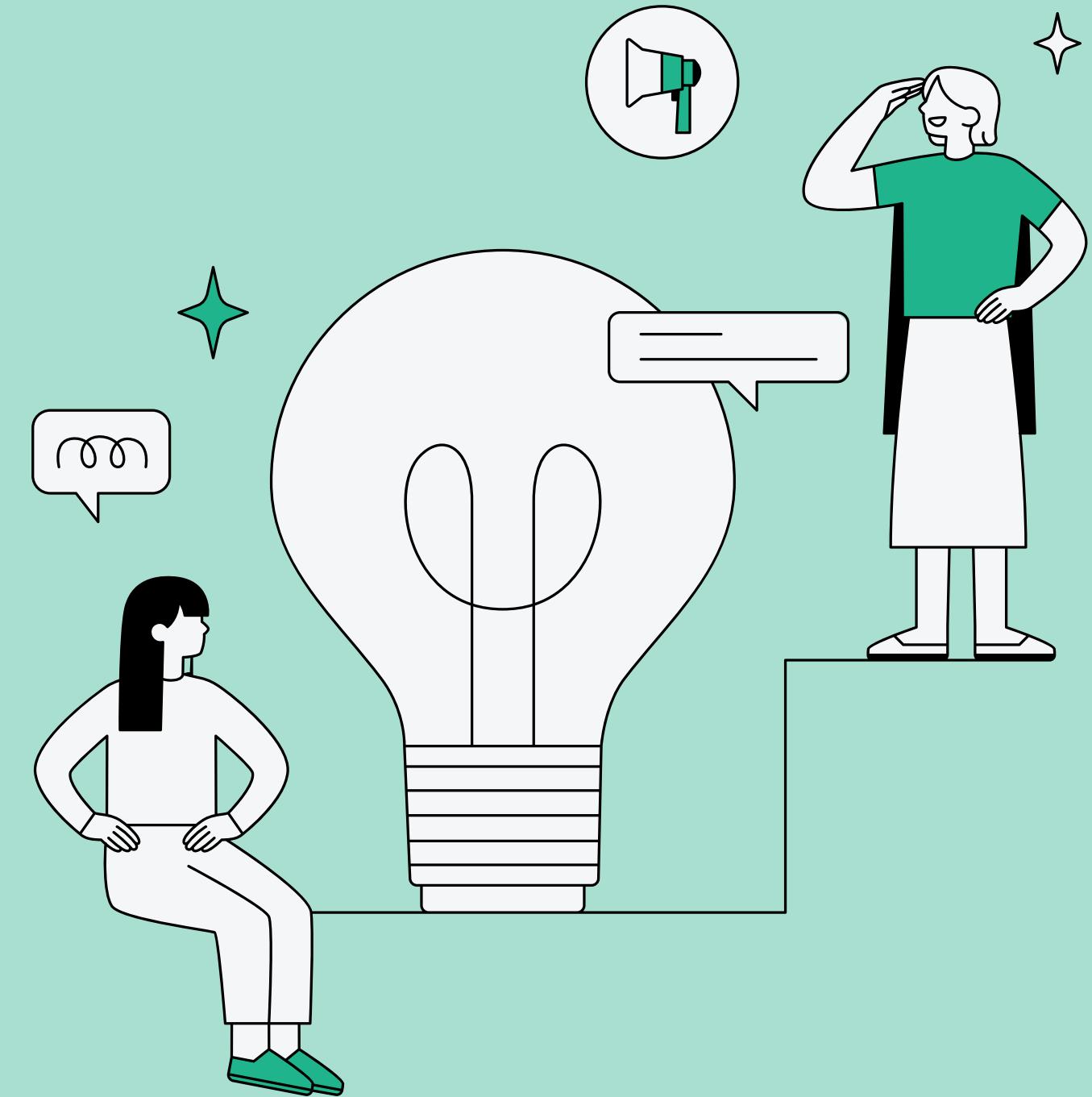
Integrating News Sentiment
and Technical Indicators



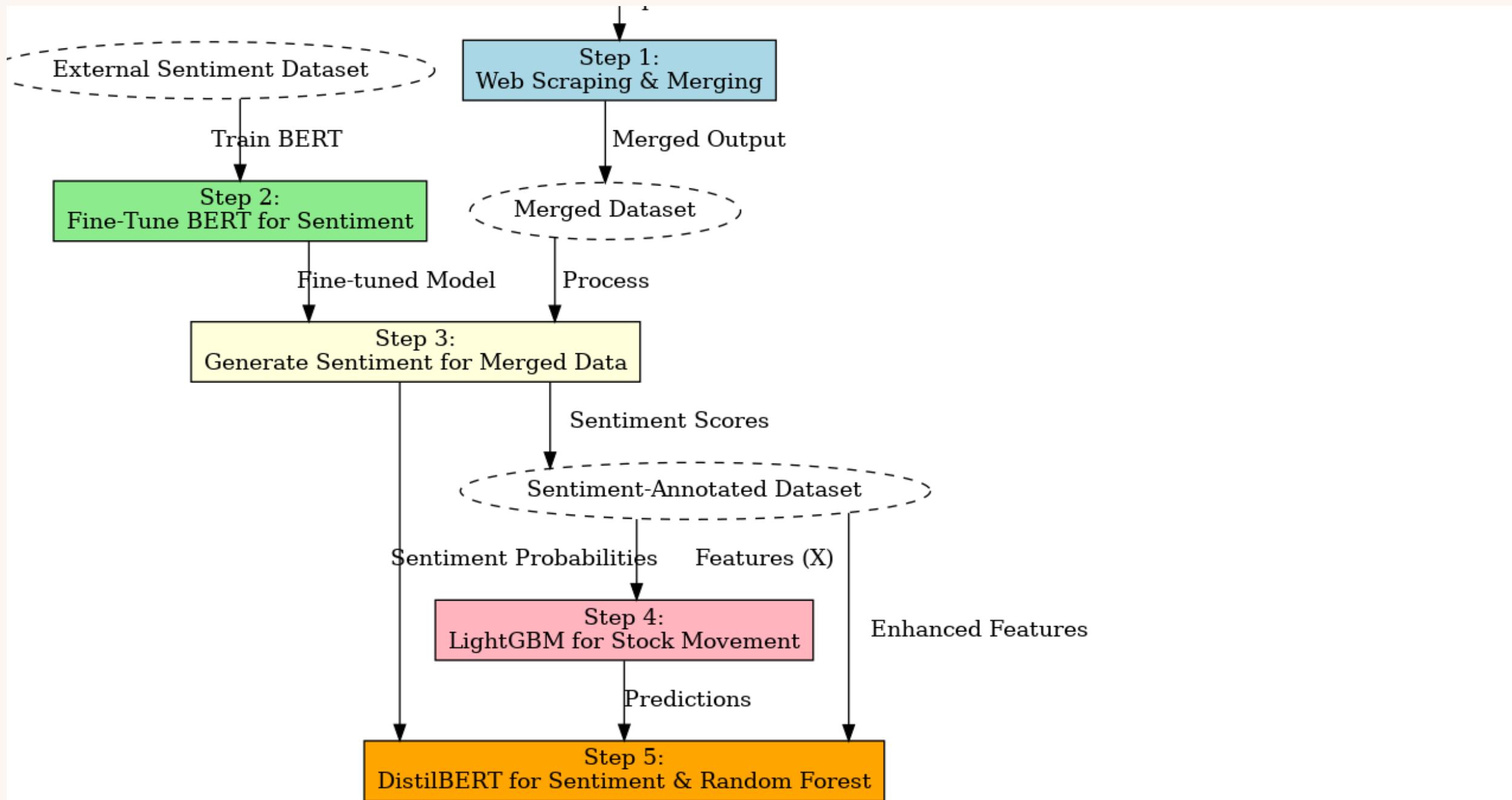
Date: 24-11-2024

Project Overview

- **Objective:** Predict stock market movement by combining:
 - Sentiment analysis of news headlines.
 - Technical stock attributes (e.g., Open, High, Low, etc.).
- **Key Steps:**
 - a. Web scraping and dataset creation.
 - b. Sentiment analysis using finetuned-BERT and DistillBERT.
 - c. Stock movement prediction with Random Forest and LightGBM.



FLOW CHART



WORKFLOW

1. Web Scraping & Merging:

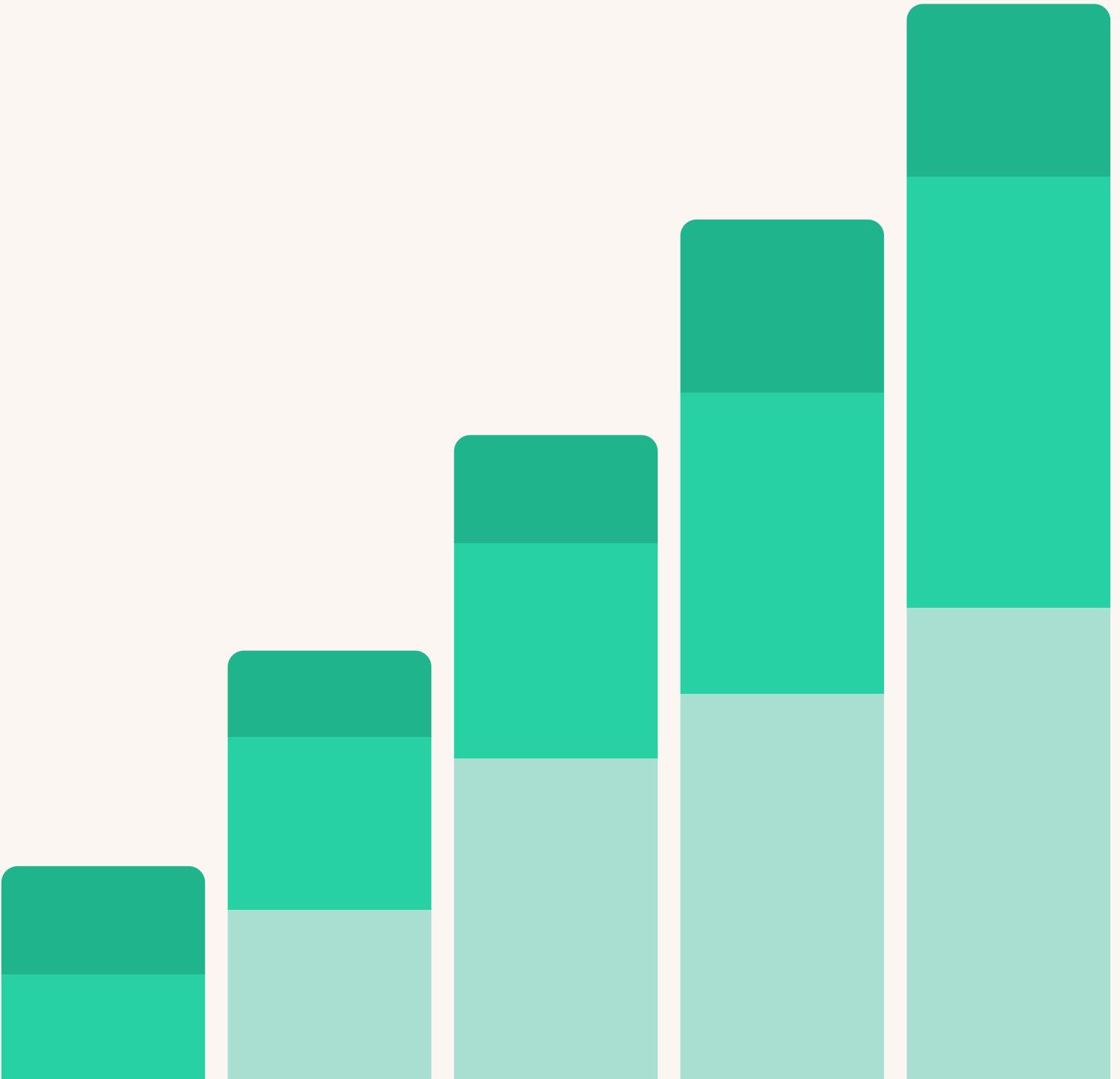
- Scrapped daily news summaries and merged with Nifty50 stock data.

2. Sentiment Analysis:

- Fine-tuned BERT for binary sentiment analysis (1/0).
- Used distill bert for sentiment prediction (negative/positive) and sentimentprobabilties.

3. Stock Prediction Models:

- Random Forest and LightGBM to predict stock movement using:
 - Sentiment probabilities.
 - Stock trading indicators.





Data and Modeling

- Features:
 - Sentiments (predicted), Stock Movements.
 - Stock indicators: Open, High, Low, Turnover, Shares Traded, sentiment probability.

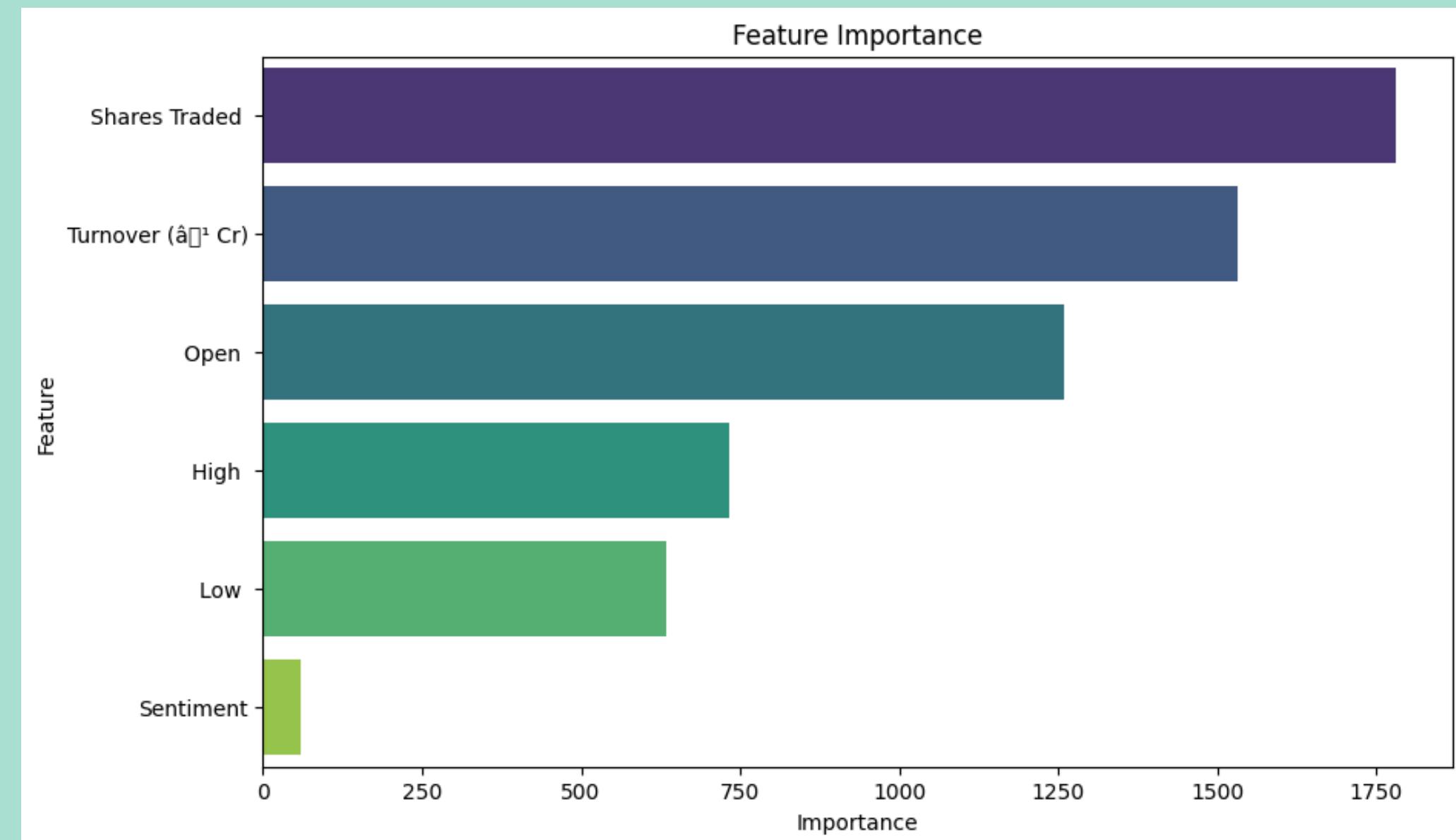
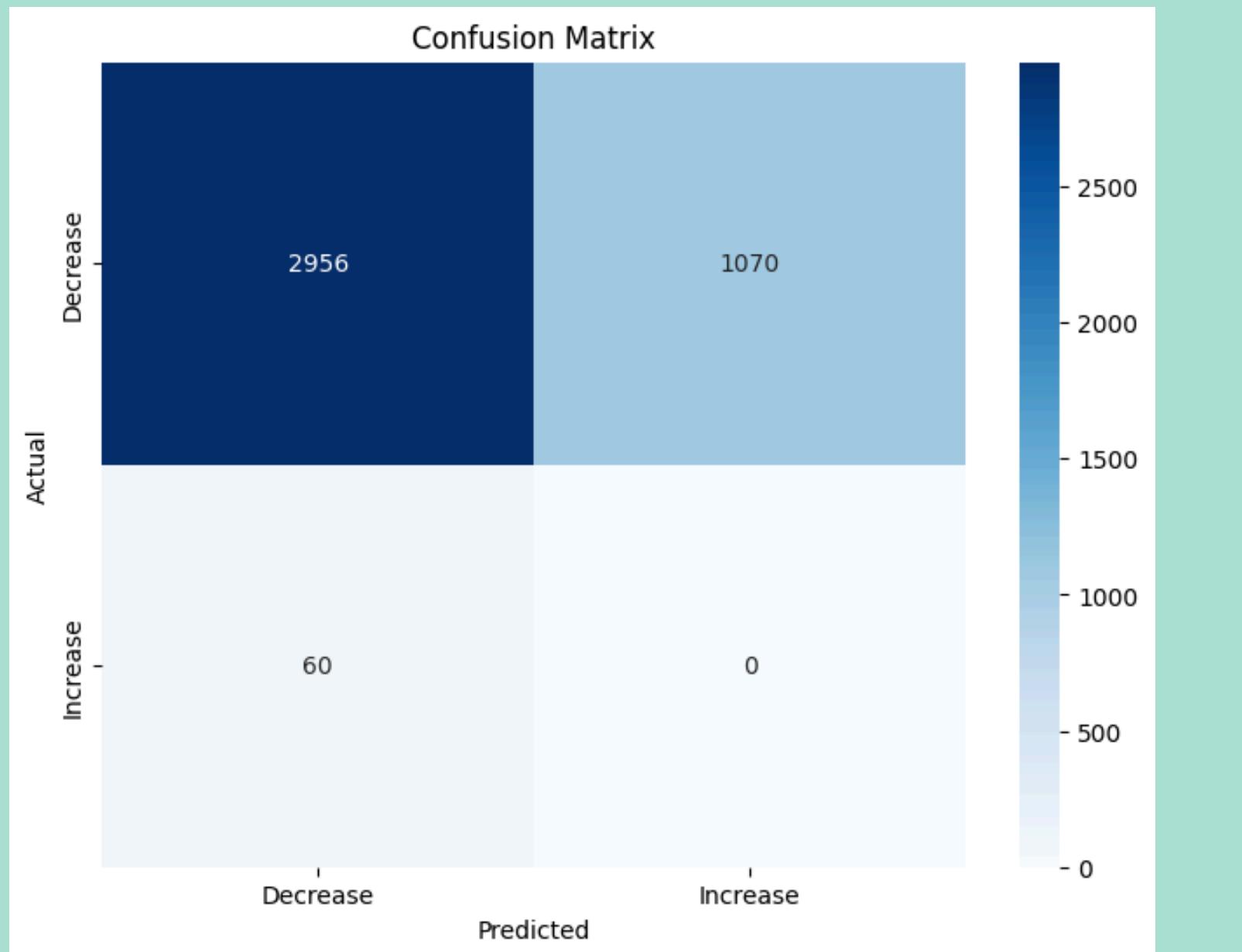
- Target Variables:
 - Sentiment classes (negative/neutral/positive).
 - Stock movement (1 = Increase, 0 = Decrease).

Tools Used:

pandas, BeautifulSoup, transformers, scikit-learn, lightgbm, Matplotlib.



key visualisations for LightGBM Model



Results for LightGBM Model

```
print("\nClassification Report:")
print(classification_report(y_test, y_pred))

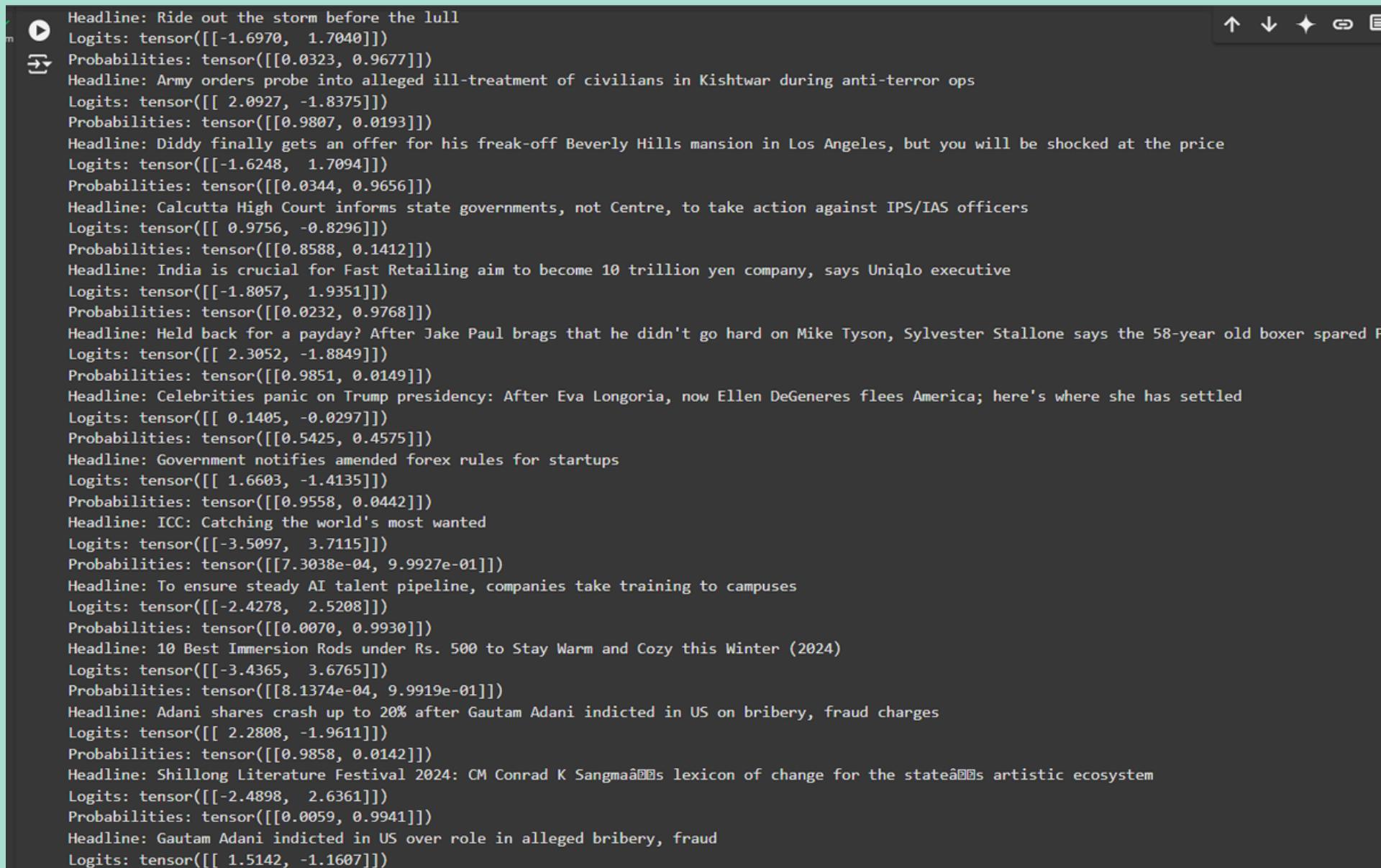
accuracy = accuracy_score(y_test, y_pred)
print(f"\nAccuracy: {accuracy:.2f}")


```

	precision	recall	f1-score	support
0	0.98	0.73	0.84	4026
1	0.00	0.00	0.00	60
accuracy			0.72	4086
macro avg	0.49	0.37	0.42	4086
weighted avg	0.97	0.72	0.83	4086

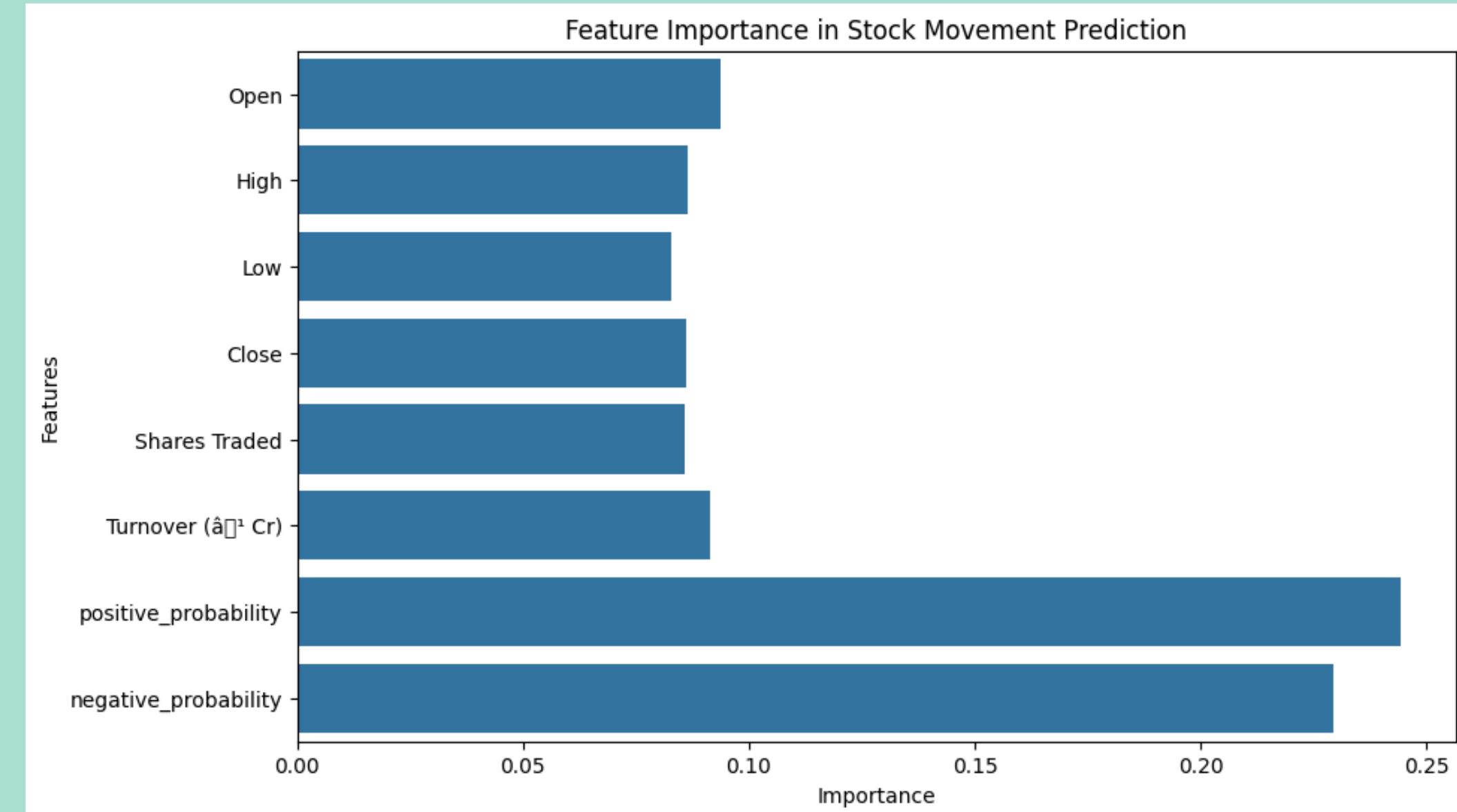
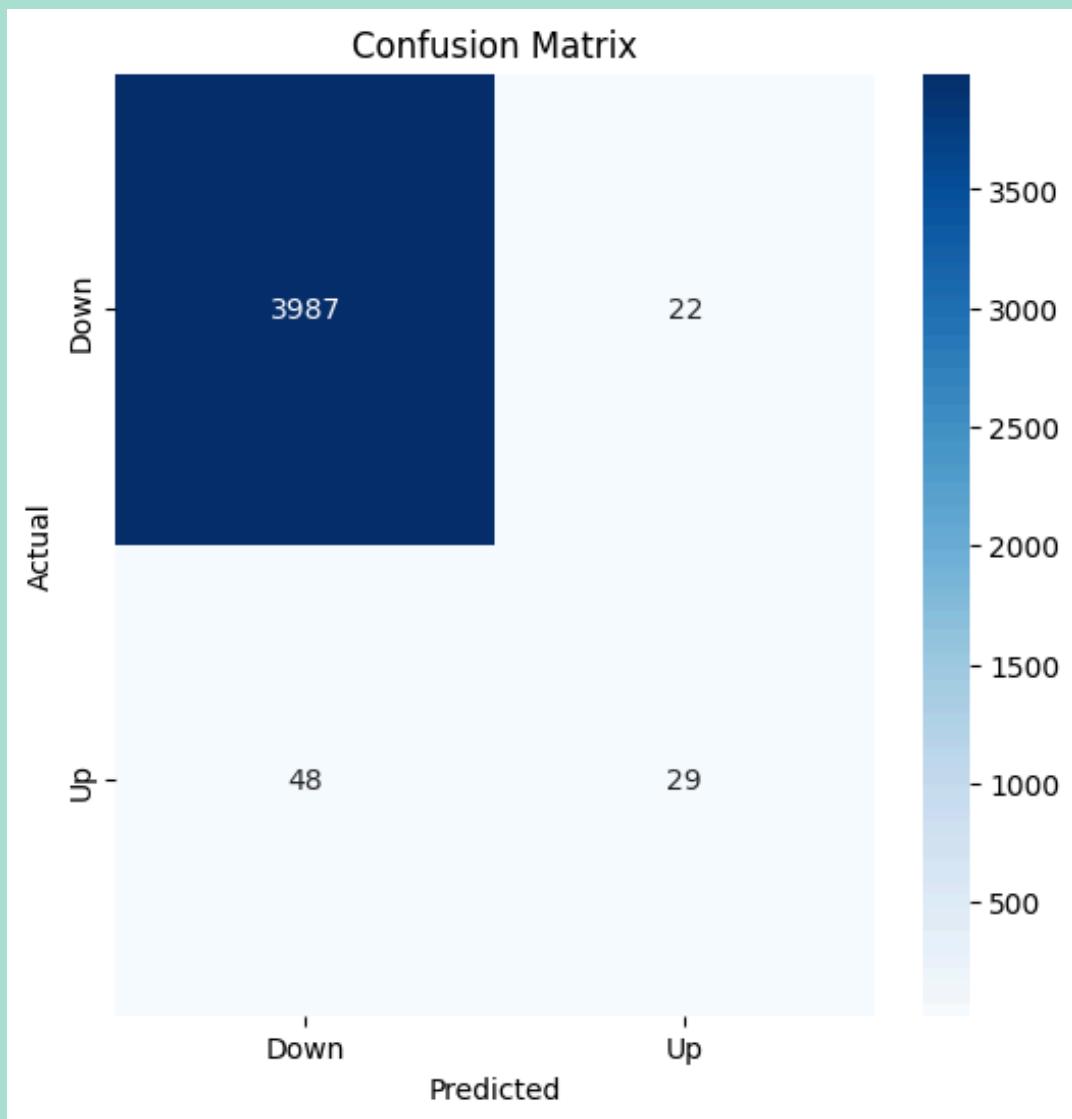
Accuracy: 0.72

PROBABILITIES AND TENSORS FROM DISTILBERT



```
m  Headline: Ride out the storm before the lull
Logits: tensor([[-1.6970,  1.7040]])
Probabilities: tensor([[0.0323, 0.9677]])
Headline: Army orders probe into alleged ill-treatment of civilians in Kishtwar during anti-terror ops
Logits: tensor([[ 2.0927, -1.8375]])
Probabilities: tensor([[0.9807, 0.0193]])
Headline: Diddy finally gets an offer for his freak-off Beverly Hills mansion in Los Angeles, but you will be shocked at the price
Logits: tensor([[-1.6248,  1.7094]])
Probabilities: tensor([[0.0344, 0.9656]])
Headline: Calcutta High Court informs state governments, not Centre, to take action against IPS/IAS officers
Logits: tensor([[ 0.9756, -0.8296]])
Probabilities: tensor([[0.8588, 0.1412]])
Headline: India is crucial for Fast Retailing aim to become 10 trillion yen company, says Uniqlo executive
Logits: tensor([-1.8057,  1.9351])
Probabilities: tensor([[0.0232, 0.9768]])
Headline: Held back for a payday? After Jake Paul brags that he didn't go hard on Mike Tyson, Sylvester Stallone says the 58-year old boxer spared Pau
Logits: tensor([[ 2.3052, -1.8849]])
Probabilities: tensor([[0.9851, 0.0149]])
Headline: Celebrities panic on Trump presidency: After Eva Longoria, now Ellen DeGeneres flees America; here's where she has settled
Logits: tensor([[ 0.1405, -0.0297]])
Probabilities: tensor([[0.5425, 0.4575]])
Headline: Government notifies amended forex rules for startups
Logits: tensor([[ 1.6603, -1.4135]])
Probabilities: tensor([[0.9558, 0.0442]])
Headline: ICC: Catching the world's most wanted
Logits: tensor([-3.5097,  3.7115])
Probabilities: tensor([[7.3038e-04, 9.9927e-01]])
Headline: To ensure steady AI talent pipeline, companies take training to campuses
Logits: tensor([[-2.4278,  2.5208]])
Probabilities: tensor([[0.0070, 0.9930]])
Headline: 10 Best Immersion Rods under Rs. 500 to Stay Warm and Cozy this Winter (2024)
Logits: tensor([-3.4365,  3.6765])
Probabilities: tensor([[8.1374e-04, 9.9919e-01]])
Headline: Adani shares crash up to 20% after Gautam Adani indicted in US on bribery, fraud charges
Logits: tensor([ 2.2808, -1.9611])
Probabilities: tensor([[0.9858, 0.0142]])
Headline: Shillong Literature Festival 2024: CM Conrad K Sangmaâ€™s lexicon of change for the stateâ€™s artistic ecosystem
Logits: tensor([-2.4898,  2.6361])
Probabilities: tensor([[0.0059, 0.9941]])
Headline: Gautam Adani indicted in US over role in alleged bribery, fraud
Logits: tensor([ 1.5142, -1.1607])
```

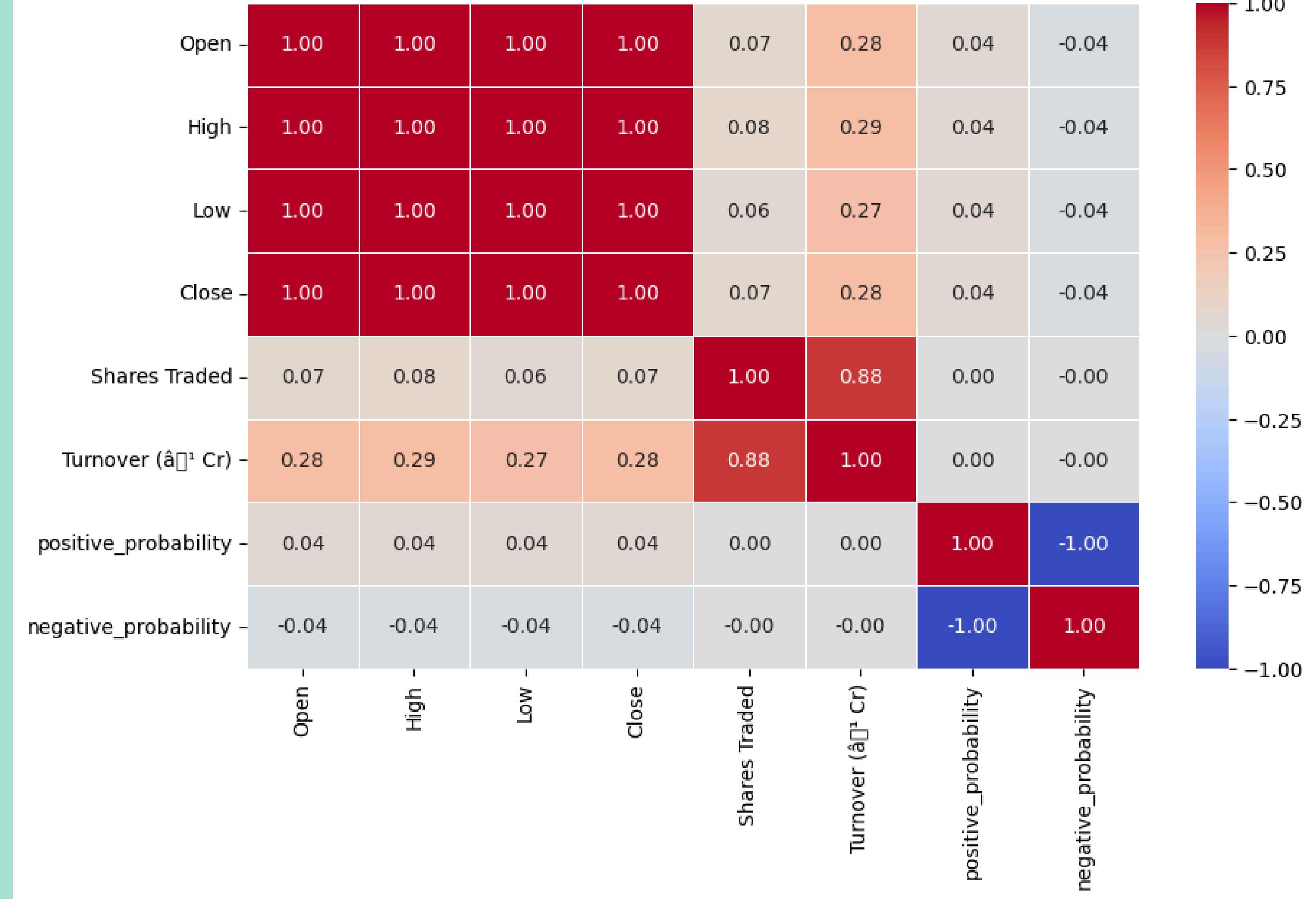
key visualisations for Randomforest with sentiment probability



Results for Randomforest with sentiment probability

Classification Report:				
	precision	recall	f1-score	support
0	0.99	0.99	0.99	4009
1	0.57	0.38	0.45	77
accuracy			0.98	4086
macro avg	0.78	0.69	0.72	4086
weighted avg	0.98	0.98	0.98	4086
Accuracy Score:				
0.982868330885952				

Correlation Heatmap



Stock Movement Prediction Models

Light BGM results

- Accuracy: 0.72%
Key Metrics:
- Precision, Recall, F1-score visualized in a classification report.

Random Forest with sentiment probability results:

- Accuracy: 0.98 %
Key Metrics:
- Precision, Recall, F1-score visualized in a classification report.

Work

Previous Work

- Initial Models: Early methods of stock market prediction relied heavily on machine learning algorithms like SVM. These models required significant feature engineering and failed to capture **sequential dependencies** inherent in time-series data
- **StockFormer**: A transformer-based model combining news sentiment (via FinBERT) and stock price data. It **outperforms** traditional models like SVMs by **improving prediction accuracy**

Ongoing Work

- **MASTER (Market-Guided Stock Transformer)**: A newer transformer model that focuses on intra-stock and inter-stock aggregation. It uses multi-head attention to capture dynamic correlations between stocks, **improving the accuracy of stock forecasting**
- Recent Developments: The latest research emphasizes multi-modal data integration. Models now incorporate news, financial metrics, and social media data to improve forecasting. This multi-source approach aims to create a more holistic prediction model

Novelty

- We initially focused on using volatile metrics such as the Volatility Index (VXN) to identify volatility patterns. However, our model's performance was suboptimal, yielding an accuracy of 55%, with recall and precision scores close to zero, indicating a significant data imbalance.
- Upon further analysis, we concluded that more data was needed to improve the model, but this added complexity. As a result, we decided to switch our approach. ([reference link](#))
- Instead of modelling with only sentiment score, we modelled it with sentiment probability scores using **DistilBERT**
- We introduce softmax layer to get **sentiment probabilities** .
- We modelled them by random forest.

Thank
you very
much!

STEMTOKLEM

