Sentiment-Driven Stock Price Prediction with Multimodal Streams

Team:
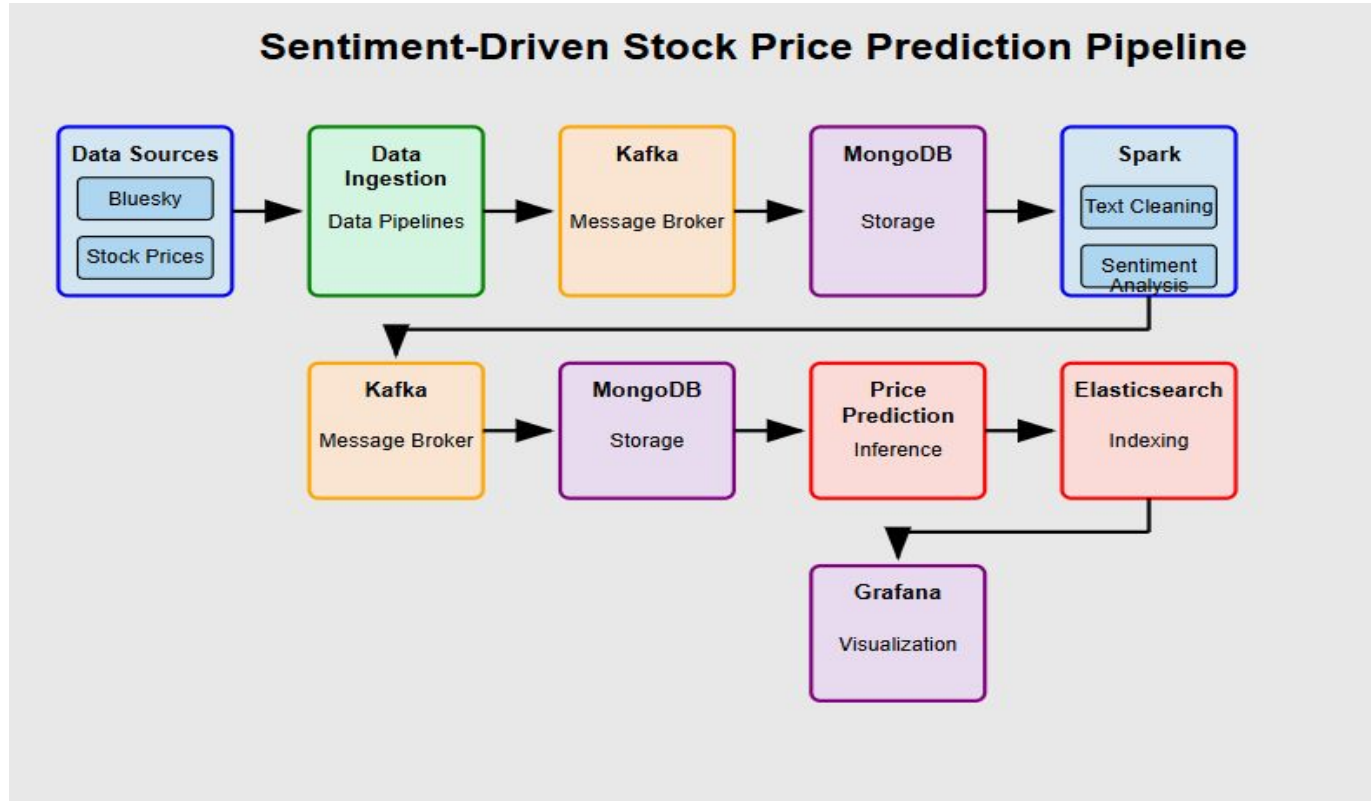
Abhipal Sharma, Harsha Mupparaju, Henry Kam, Nived Damodaran

04/01/2025

Problem Statement

Stock prices are heavily influenced by market sentiment, which can be gleaned from social media discussions (e.g., Twitter, Reddit), financial news, and other public discourse. Traditional stock market analysis focuses on historical prices and technical indicators, but real-time sentiment analysis has the potential to predict price movements and market reactions more accurately. This project addresses the challenge of integrating multimodal data streams and performing real-time sentiment analysis at scale.

NYU

System Architecture

Data Sources

1. YFinance library
   a. Gathers stock price details of the companies, such as the Closing Price, Opening Price, Low Price, High Price and Volume traded for each day.
2. Bluesky API
   a. Official API of Bluesky
   b. Gathers data such as the text content, date posted, number of likes, number of replies, number of quotes, number of reposts.

Message Broker - Kafka

- Kafka enables real-time streaming of data.
- Has a high throughput - can handle millions of messages per second.
- It is scalable.
- Allows decouples architecture where the producers and consumers don't need to know about each other.

```python
message = {
    "symbol": symbol,
    "closing_price": round(float(row["Close"]), 2),
    "opening_price": round(float(row["Open"]), 2),
    "high": round(float(row["High"]), 2),
    "low": round(float(row["Low"]), 2),
    "volume": round(float(row["Volume"]), 2),
    "timestamp": row.name.strftime("%Y-%m-%d")
}
```
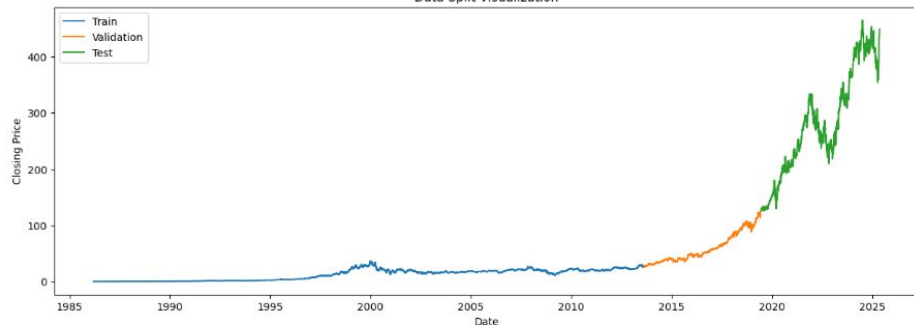
```python
bluesky_message = {
    "id": post.get("cid"),
    "company": company,
    "timestamp": record.get("createdAt", "").split("T")[0] if "createdAt" in record else "unknown",
    "text": record.get("text", ""),
    "replyCount": post.get("replyCount", 0),
    "likeCount": post.get("likeCount", 0),
    "repostCount": post.get("repostCount", 0),
    "quoteCount": post.get("quoteCount", 0)
}
```

NYU

Sentiment Analysis and Price Prediction

- FinBERT is a domain-specific language model based on BERT, fine-tuned specifically for financial text. It is designed to understand sentiment and context in financial documents.
- Predicts positive, negative, or neutral sentiment for a sentence or paragraph.
- We curate the data by merging the sentiment scores of the posts with the OHLCV data of the stocks, and use a custom model to predict the prices.
- The custom model is a Long Short Term Memory (LSTM) model with 3 linear layers that takes the average sentiment scores of the company for each day and the stock data and tries to predict the price for the next day.
- The price prediction model achieves good accuracy as can be seen from the plots on the next slide. Although the next day prices are not exactly predicted, but the trend of the graph is predicted very well.
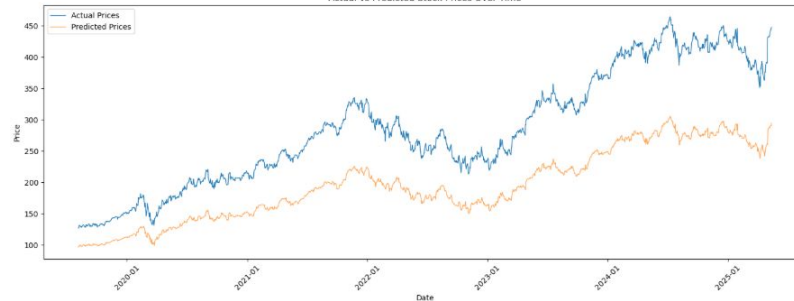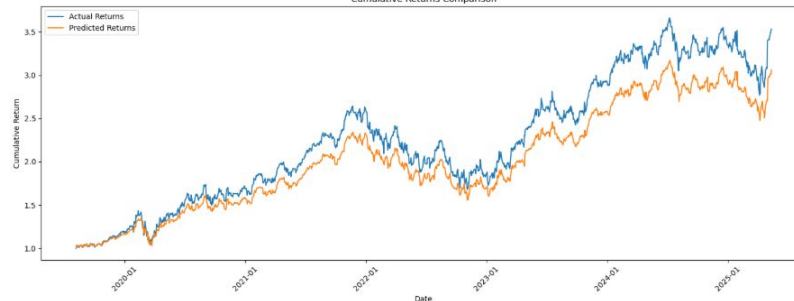
NYU

Plots

Epoch [1/5], Train Loss: 0.081291, Val Loss: 3.636621
Epoch [2/5], Train Loss: 0.027421, Val Loss: 2.130667
Epoch [3/5], Train Loss: 0.011747, Val Loss: 1.344507
Epoch [4/5], Train Loss: 0.009740, Val Loss: 0.847174
Epoch [5/5], Train Loss: 0.007637, Val Loss: 0.397124

MongoDB - Temporary Store

- Store intermediate data from publishers
- Updates data with sentiment analysis
- Aggregates tweets and prices from stream to create informed inference

2025-05-15 14:12:49,680 [INFO] Received message from topic twitter_raw: {'id': 'bafyreif3jvnesctl
2025-05-15 14:12:49,681 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,681 [INFO] Received message from topic twitter_raw: {'id': 'bafyreia334t2dqyu
2025-05-15 14:12:49,682 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,682 [INFO] Received message from topic twitter_raw: {'id': 'bafyreiazwbmuh6pc
2025-05-15 14:12:49,683 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,683 [INFO] Received message from topic twitter_raw: {'id': 'bafyreiedg7gekv3w
2025-05-15 14:12:49,684 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,684 [INFO] Received message from topic twitter_raw: {'id': 'bafyreiaimsd2yhwc
2025-05-15 14:12:49,685 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,685 [INFO] Received message from topic twitter_raw: {'id': 'bafyreiedgyofwv3t
2025-05-15 14:12:49,685 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,685 [INFO] Received message from topic twitter_raw: {'id': 'bafyreiciwuvnvi72
2025-05-15 14:12:49,686 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,686 [INFO] Received message from topic twitter_raw: {'id': 'bafyreifofl3cm5kb
2025-05-15 14:12:49,687 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,687 [INFO] Received message from topic twitter_raw: {'id': 'bafyreicsju6emtxp
2025-05-15 14:12:49,688 [INFO] Inserted message into MongoDB: microsoft, twitter_raw
2025-05-15 14:12:49,688 [INFO] Received message from topic twitter_raw: {'id': 'bafyreihx3larvio2
2025-05-15 14:12:49,689 [INFO] Inserted message into MongoDB: microsoft, twitter_raw

NYU

Elasticsearch

```
(sentiment-stocks) abhipalsharma@Wolfe-MacBook-Air storage % curl -s "http://localhost:9200/sentiment-index/_count?pretty"
{
  "count" : 12867,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  }
}
(sentiment-stocks) abhipalsharma@Wolfe-MacBook-Air storage % curl -s "http://localhost:9200/sentiment-index/_mapping?pretty"
{
  "sentiment-index" : {
    "mappings" : {
      "properties" : {
        "average_sentiment" : {
          "type" : "long"
        },
        "close" : {
          "type" : "float"
        },
        "high" : {
          "type" : "float"
        },
        "low" : {
          "type" : "float"
        },
        "open" : {
          "type" : "float"
        },
        "predicted_price" : {
          "type" : "float"
        },
        "target_price" : {
          "type" : "float"
        },
        "ticker" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
        "timestamp" : {
          "type" : "date"
        },
        "volume" : {
          "type" : "long"
        }
      }
    }
  }
}
```

```
"hits" : {
  "total" : {
    "value" : 10000,
    "relation" : "gte"
  },
  "max_score" : 1.0,
  "hits" : [
    {
      "_index" : "sentiment-index",
      "_id" : "7ehN1JYBntpnBzr1_7MG",
      "_score" : 1.0,
      "_source" : {
        "close" : 237.6036376953,
        "high" : 240.6661183041,
        "low" : 237.1869153202,
        "open" : 239.9295785121,
        "volume" : 21863100,
        "ticker" : "MSFT",
        "timestamp" : "2021-05-21T00:00:00Z",
        "average_sentiment" : -1,
        "target_price" : 240.1427831204,
        "predicted_price" : 240.7324899221
      }
    },
    {
      "_index" : "sentiment-index",
      "_id" : "7uhN1JYBntpnBzr1_7MG",
      "_score" : 1.0,
      "_source" : {
        "close" : 243.040512085,
        "high" : 243.4087893868,
        "low" : 239.8714255678,
        "open" : 240.1427831204,
        "volume" : 21411500,
        "ticker" : "MSFT",
        "timestamp" : "2021-05-24T00:00:00Z",
        "average_sentiment" : 0,
        "target_price" : 243.9999827054,
        "predicted_price" : 250.1417338127
      }
    },
    {
      "_index" : "sentiment-index",
      "_id" : "7-hN1JYBntpnBzr1_7MG",
      "_score" : 1.0,
      "_source" : {
        "close" : 243.9515228271,
        "high" : 244.9497342108,
        "low" : 243.0793041694,
        "open" : 243.9999827054,
        "volume" : 17704300,
        "ticker" : "MSFT",
        "timestamp" : "2021-05-25T00:00:00Z",
        "average_sentiment" : -1,
        "target_price" : 243.670415945,
        "predicted_price" : 241.6684960821
      }
    }
```

Sentiment Analysis

```json
{
  "_id": "68263da68209c582b33cad76",
  "source_topic": "twitter_raw",
  "kafka_timestamp_ms": 1747336613947,
  "symbol": "microsoft",
  "timestamp": "2025-05-15",
  "data": {
    "type": "sentiment",
    "text": "The current versions of Windows & Microsoft are the worst I've ever experienced. There is nothing efficient about the way they're running things rn.",
    "id": "bafyreibwu55dcjt5bxz3zlswvrcuxbbziyo7ro46qod6beoqg6bs7lnawy",
    "original_text": "The current versions of Windows & Microsoft are the worst I've ever experienced.\n There is nothing efficient about the way they're running things rn.",
    "processed_at": "2025-05-15T19:17:21.221953",
    "sentiment_label": "negative",
    "sentiment_score": 0.0047799897193909
  }
}
```

Grafana

DEMO!