Sentiment-Driven Stock Price Prediction with Multimodal Streams

Sai Harsha Mupparaju
sm12754@nyu.edu

Nived Damodaran
nd2746@nyu.edu

Henry Kam (Online Course)
hjk9412@nyu.edu

Abhipal Sharma
as20410@nyu.edu

***Index Terms*—Stock Price Prediction, Sentiment Analysis, Multimodal Data, Real-Time Data Processing, Machine Learning, Financial News Analysis, Natural Language Processing (NLP), Time Series Forecasting, Big-Data**

I. Introduction

The goal of this project is to build a **real-time sentiment analysis** system for the stock market using multimodal data streams. This system will aggregate data from various sources such as Twitter, Reddit (r/stocks), financial news, and historical stock prices to perform sentiment analysis and assess its impact on stock price movements. The model uses sentiment analysis (with Spark ML or BERT) to analyze news and social media data, helping predict stock trends more accurately.

II. Objective

- Ingest **real-time data** from Twitter, Reddit, and financial news
- Perform **sentiment analysis** using **Spark ML** or **BERT**
- Correlate **sentiment trends** with **price movements**
- Store processed data in **HDFS/MongoDB** and **Elasticsearch**
- Visualize insights via **interactive dashboards**
- Generate **predictive signals** for stock movements

III. 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.

IV. Data Sources

The following data sources will be used in this project:

- **Twitter**: Tweets related to stock tickers (e.g. AAPL, TSLA, GME) will be streamed in real-time using the Twitter API.
- **Reddit**: Posts and comments from subreddits like r/stocks, r/Investing will be ingested using Pushshift or PRAW.
- **Financial News**: Articles, headlines, and metadata from financial news outlets using NewsAPI.
- **Historical Stock Price Data**: Stock price data (OHLCV: Open, High, Low, Close, Volume) using Alpha Vantage, etc will be stored in HDFS/Hive.

V. Proposed Architecture

- **Data Ingestion** (Kafka)
  - **Producers**: Twitter (tweets), Reddit (posts), News (headlines), Stock Prices
- **Data Processing** (Spark Structured Streaming)
  - Clean text, extract tickers, classify sentiment
  - Store processed data in **HDFS**
- **Sentiment Classification**
  - **Spark ML** (Logistic Regression/Naive Bayes) for real-time
  - **Fine-tuned BERT/FinBERT** for batch (accuracy)
- **Storage & Indexing**
  - **HDFS/MongoDB**
  - **Elasticsearch**: Indexed for fast queries
- **Visualization**
  - **Grafana**: To visualize the insights from the models
  - **Gradio/Notebook**: Query interface
- **Prediction**
  - Correlate sentiment with prices
  - Predict using **XGBoost/LSTM/ARIMA**

VI. Expected Outcomes

- **Real-time sentiment scores** for stocks from social media and news
- **Visualized sentiment trends** correlated with market movements
- **Predictive signals** for stock price changes based on sentiment

VII. Conclusion

This project develops a stock prediction tool that **combines sentiment analysis** from news and social media with **historical price patterns**. The system identifies how market sentiment **correlates with price movements**, providing traders with **data-driven insights** beyond traditional analysis. By automating this process, the tool offers a practical way to enhance trading decisions.