Research Report

Title: Advanced Data Correlation and Enrichment Using LSTM

Author: Shriyansh Thote

1. Introduction

The exponential growth in financial data, especially in stock markets, has intensified the need for intelligent models that can detect deep patterns and future trends. This research explores the application of Long Short-Term Memory (LSTM) networks to detect S-Curve growth points, a critical phase in a stock's lifecycle where it transitions from slow growth to rapid acceleration. The work integrates log analysis, technical indicators, and market data, forming the initial foundation of Multimodal AI research.

2. Objectives

- To analyze and correlate long-term historical stock data to detect early signals of significant price growth.

- To use LSTM models to learn temporal patterns from pre-2020 stock data and predict S-Curve Start Points in post-2020 stocks.

- To integrate fundamental indicators (e.g., P/E Ratio, EPS) and technical signals into the data enrichment process.

- To enable sector-wise insights into market behavior using machine learning.

3. Approach & Methodology

3.1 Data Collection and Preparation

- Historical stock data was collected using the `yfinance` library.

- IPO dates were verified using the NSE website to avoid including pre-listing data.

- Stocks were split into: Pre-2020 stocks for training and Post-2020 stocks for prediction.

3.2 Feature Engineering

- Yearly average prices, volatility, growth rate, and intrinsic value were computed.

- Sector classification was added for contextual analysis.

3.3 LSTM Model Architecture

- Two-layer LSTM with dropout layers.

- Dual-output architecture: price\_output → Predicts expected S-Curve price; year\_output → Predicts years to S-Curve start.

3.4 Sequence Formatting

- A sliding window approach was used to create 5-year input sequences.

- Data was normalized using MinMaxScaler to preserve trend directionality.

4. Techniques for S-Curve Detection

Method 1: Local Minima Detection

- Identified historical local minima.

- Checked for a ≥10% price rise for the next 3 years.

- Labeled such points as S-Curve starts.

Method 2: Global Minimum-Based Analysis

- Found global minima (excluding listing price).

- If followed by strong upward movement, marked as S-Curve start.

Method 3: Rolling Growth Analysis

- Used 5-year rolling windows.

- Criteria: ≥15% average growth, with 60% of years showing strong returns.

5. Results & Observations

Model Accuracy

- Simple LSTM model achieved ~80–90% accuracy.

- Bidirectional LSTM (Enhanced LSTM) achieved up to 99.43% accuracy in certain cases.

Prediction Behavior

- Method 1 and 2 yielded mostly 0s and 1s (limited resolution).

- Method 3 produced a well-distributed range of predicted values, aligning with real-world expectations.

Sector-Wise Analysis

- Stocks were grouped into sectors (e.g., Technology, Financials).

- Identified “hot sectors” with early S-Curve signals.

- Within sectors, identified underperformers lagging their peers.

6. Challenges Faced

- IPO Date Inaccuracy: `yfinance` provided unreliable pre-IPO data, affecting early models. Resolved by scraping accurate IPO data from NSE.

- Sector Label Gaps: Some stocks had no sector info via API. Grouped them under 'Unknown' for consistency.

- Model Overfitting: LSTM needed regularization to avoid overfitting on short-term noise.

- Prediction Range Compression: Early models predicted narrow ranges (mostly 0 or 1). Improved via rolling-window growth analysis (Method 3).

7. Tools & Technologies Used

- Python (NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow/Keras)

- yfinance, NSE website scraping

- LSTM & Bidirectional LSTM neural networks

- Data normalization, outlier removal, and sequence slicing

8. Future Work

- Incorporate global market factors to model external influence.

- Develop sector-specific LSTM models for finer predictions.

- Integrate real-time data feeds for live trend detection.

- Extend to multi-modal models combining logs, prices, and financial news for deeper insight.

9. Conclusion

This research successfully applies LSTM architectures for advanced stock data correlation and predictive enrichment. The methods developed — particularly the rolling-growth detection strategy — provide robust tools for forecasting S-Curve points. By combining temporal deep learning with financial logic, this work sets a foundation for intelligent, sector-aware market prediction systems.