# Biocon FDA Process & Stock Price Correlation Analysis

## 4-Day Plan of Action (POA)

### Project Overview

Develop an AI-driven system to analyze the correlation between FDA drug approval processes, news sentiment, and stock price movements for Biocon, with initial focus on Semglee® (insulin glargine-yfgn). The system will use machine learning models including LSTM and deep learning for predictive analysis.

## Day 1: Data Collection & Infrastructure Setup

### Morning (4 hours)

**Stock Price Data Collection**

- **NSE/BSE Data Sources:**
  - Download Biocon (BIOCON.NS) daily stock prices (Jan 1, 2015 - June 20, 2025)
  - Collect OHLCV data, trading volumes, market cap changes
  - Sources: NSE official API, Yahoo Finance, Alpha Vantage

**Market Benchmark Data**

- **Nifty 50 Index:** Daily prices for market correlation analysis
- **Nifty Pharma Index:** Sector-specific benchmark for pharmaceutical industry
- **Peer Companies:** Similar pharma companies for comparative analysis

### Afternoon (4 hours)

**News Data Collection**

- **Primary Sources:**
  - Bloomberg Terminal/API for financial news
  - Reuters pharmaceutical sector news
  - Pharmaceutical industry publications (BioPharma Dive, FiercePharma)
  - FDA official announcements and press releases

**Semglee®-Specific Data**

- **FDA Process Timeline:**
  - Initial application dates
  - Clinical trial announcements
  - FDA meeting schedules
  - Approval milestones
  - Post-market surveillance updates
- **News Categories to Track:**
  - FDA submissions and responses

  - Clinical trial results
  - Regulatory approvals/rejections
  - Patent updates
  - Partnership announcements
  - Competitor news

### Data Storage Setup

- Set up PostgreSQL database with tables for:
  - Stock prices (daily/intraday)
  - News articles with timestamps
  - FDA process milestones
  - Market indices
  - Sentiment scores

# Day 2: Model Development & Training

## Morning (4 hours)

#### Data Preprocessing

- **Stock Data Cleaning:**
  - Handle missing values, stock splits, dividends
  - Calculate technical indicators (RSI, MACD, Bollinger Bands)
  - Normalize price data for model training
- **News Data Processing:**
  - Text cleaning and tokenization
  - Remove duplicates and irrelevant content
  - Create news-stock price alignment based on timestamps

#### Feature Engineering

- **Market Features:**
  - Price momentum indicators
  - Volume-weighted average prices
  - Market beta calculations
  - Sector performance relative metrics
- **News Features:**
  - Sentiment scores (VADER, TextBlob, custom pharma lexicon)
  - News frequency and intensity metrics
  - FDA-specific keyword identification
  - News source credibility weighting

## Afternoon (4 hours)

#### Model Architecture Development

- **LSTM Model for Time Series:**
  - Multi-layered LSTM for stock price prediction
  - Input features: historical prices, volume, market indices, sentiment scores
  - Output: Next-day price movement probability
- **Sentiment Analysis Model:**
  - Fine-tuned BERT model for pharmaceutical news

    - Custom vocabulary for FDA process terminology
    - Binary/multi-class classification for news impact
- **Ensemble Model:**
    - Combine LSTM predictions with sentiment analysis
    - Random Forest for feature importance analysis
    - XGBoost for non-linear relationship detection

#### Training Process

- **Data Split:** 70% training (2015-2021), 15% validation (2022-2023), 15% test (2024-2025)
- **Cross-validation:** Time series split to prevent data leakage
- **Hyperparameter tuning:** Grid search with walk-forward analysis

## Day 3: Model Testing & Debugging

### Morning (4 hours)

#### Model Evaluation

- **Performance Metrics:**
    - RMSE and MAE for price predictions
    - Precision, Recall, F1-score for sentiment classification
    - Sharpe ratio for trading strategy performance
    - Maximum drawdown analysis
- **Statistical Significance Testing:**
    - T-tests for news impact on stock returns
    - Correlation analysis between sentiment and price movements
    - Granger causality tests for news-price relationships

#### Debugging & Optimization

- **Model Diagnostics:**
    - Learning curves analysis
    - Feature importance visualization
    - Residual analysis for prediction errors
    - Overfitting detection and mitigation

### Afternoon (4 hours)

#### Control for Market Effects

- **Market Neutralization:**
    - Calculate Biocon beta against Nifty 50 and Nifty Pharma
    - Implement market-adjusted returns
    - Sector rotation effects analysis
- **Event Study Methodology:**
    - Define event windows around FDA announcements
    - Calculate abnormal returns using market model
    - Statistical significance testing for event impact

#### Algorithm Refinement

- **Feature Selection:**
    - Remove multicollinear features

    - Implement LASSO regularization
    - Use mutual information for feature ranking
- **Model Ensemble:**
    - Weighted voting based on historical performance
    - Dynamic model selection based on market conditions
    - Confidence intervals for predictions

# Day 4: Validation & Future Framework Setup

## Morning (4 hours)

### New Drug Testing

- **Select Test Case:** Another Biocon drug or recent FDA submission
- **Data Collection:** Apply same methodology to new drug
- **Model Application:** Test trained models on new dataset
- **Performance Comparison:** Validate model generalizability

### Accuracy Assessment

- **Backtesting Results:**
    - Out-of-sample performance metrics
    - Comparison with buy-and-hold strategy
    - Risk-adjusted returns analysis
    - Transaction cost considerations

## Afternoon (4 hours)

### Scalable Framework Development

- **Automated Pipeline:**
    - Real-time news scraping and processing
    - Automatic sentiment scoring
    - Daily model predictions and updates
    - Alert system for significant events
- **User Interface Design:**
    - Input system for drug/company name
    - Automated data collection triggers
    - Real-time dashboard for monitoring
    - Historical analysis reports

### Documentation & Deployment

- **Technical Documentation:**
    - Model architecture specifications
    - Data pipeline documentation
    - API endpoint definitions
    - Performance benchmarks

# Key Technical Components

## Data Sources Integration

```
python
# Primary data sources
STOCK_APIS = ['NSE', 'Yahoo Finance', 'Alpha Vantage']
NEWS_SOURCES = ['Bloomberg', 'Reuters', 'FDA.gov', 'BioPharma Dive']
BENCHMARK_INDICES = ['NIFTY50', 'NIFTYPHARMA']
```

## Model Stack

- **Deep Learning:** TensorFlow/PyTorch for LSTM implementation
- **NLP Processing:** Hugging Face Transformers, spaCy
- **Time Series:** statsmodels, Prophet for trend analysis
- **Machine Learning:** scikit-learn, XGBoost, LightGBM

## Expected Deliverables

1. **Trained AI Model** with 70%+ accuracy in predicting price direction
2. **Automated News Impact Scoring** system
3. **Real-time Monitoring Dashboard**
4. **Scalable Framework** for any pharmaceutical company/drug
5. **Comprehensive Performance Report** with statistical significance tests

## Risk Considerations

- **Data Quality:** Ensure news timestamp accuracy
- **Market Noise:** Control for broader market movements
- **Regulatory Changes:** Account for evolving FDA processes
- **Insider Information:** Identify and flag unusual trading patterns

## Future Enhancements

- **Multi-company Analysis:** Extend to entire pharmaceutical sector
- **Real-time Alerts:** Push notifications for significant events
- **Portfolio Optimization:** Integration with trading strategies
- **Regulatory Intelligence:** Automated FDA filing monitoring

**Success Metrics:**

- Model accuracy >70% for price direction prediction
- Statistically significant correlation between news sentiment and stock returns
- Successful validation on new drug dataset
- Scalable framework ready for production deployment