1 FinTech Data Curator - Design Justification Document

CS4063 - Natural Language Processing Assignment 1
Author: M Shuja Uddin 22i2553 **Date:** September 18, 2025

1.1 Executive Summary

This document justifies the design decisions made in developing the FinTech Data Curator, a Python system that collects minimal feature sets for predicting next-day stock and cryptocurrency prices. The system successfully combines structured numerical data with unstructured textual information to create comprehensive datasets suitable for financial prediction models.

1.2 Feature Selection Rationale

1.2.1 Structured Data Features

1. Core Price Data (OHLCV) - **Open, High, Low, Close, Volume**: Essential baseline features representing market activity and price movements - **Rationale**: These form the foundation of technical analysis and are universally available across all financial instruments - **Predictive Value**: Price patterns and volume trends are fundamental indicators of market sentiment and momentum

2. Technical Indicators - **Moving Averages (5, 10, 20 days)**: Capture short, medium, and longer-term price trends - **RSI (Relative Strength Index)**: Identifies overbought/oversold conditions indicating potential reversals - **Bollinger Bands**: Measure volatility and price extremes relative to historical norms - **Daily Returns & Volatility**: Quantify price change magnitude and market uncertainty - **Phase 1 Enhancement - Advanced Indicators**: - **MACD (Moving Average Convergence Divergence)**: Trend-following momentum indicator with signal line and histogram - **Stochastic Oscillator (%K, %D)**: Momentum indicator comparing closing price to price range over time - **Williams %R**: Momentum indicator measuring overbought/oversold levels - **Phase 1 Enhancement - Market Context**: - **VIX (Fear Index)**: Market volatility and investor sentiment indicator - **DXY (Dollar Index)**: US Dollar strength relative to major currencies - **Treasury 10Y**: 10-year Treasury yield indicating interest rate environment - **S&P 500 Correlation**: Market-wide performance correlation

Justification: These indicators are: - Computationally efficient and widely used in quantitative finance - Complementary in capturing different aspects of price behavior (trend, momentum, volatility, market context) - Proven effective in academic literature for short-term price prediction - **Phase 1 Enhancement**: Advanced indicators (MACD, Stochastic, Williams %R) provide sophisticated momentum analysis - **Market Context**: VIX, DXY, and Treasury rates add macro-economic perspective for comprehensive market understanding

1.2.2 Unstructured Data Features

1. News Headlines & Summaries - **Source Selection**: 16+ comprehensive RSS feeds including: - **Financial News**: Yahoo Finance, MarketWatch, Reuters, Bloomberg, CNBC, Seeking Alpha, Benzinga, Financial Times, TheStreet, Fool - **Cryptocurrency**: CoinDesk, CoinTelegraph - **Phase 1 Enhancement - Regulatory Sources**: SEC Press Releases, Federal Reserve announcements, Treasury Department, CFTC, FINRA - **Rationale**: News events significantly impact short-term price movements, especially earnings announcements, regulatory changes, and market sentiment shifts. Multiple sources ensure diverse coverage and reduced bias. Regulatory sources provide early signals for policy-driven market movements.

2. Sentiment Analysis - **Implementation**: TextBlob-based sentiment scoring (0-1 scale) with enhanced relevance filtering - **Justification**: Market psychology drives price movements; sentiment provides emotional context missing from pure technical analysis. Improved filtering ensures higher quality sentiment signals.

3. Relevance Scoring - **Algorithm**: Multi-tier relevance scoring with symbol mentions (0.9), company names (0.8), financial keywords (0.6), sector terms (0.4), and general business (0.3) - **Purpose**: Filter noise and focus on news directly impacting the target asset while maintaining contextual market information

---

1.3 MINIMALITY ARGUMENT

1.3.1 Why This Feature Set is Sufficient

1. Comprehensive Market Aspects - **Price Action**: OHLCV data captures all transaction information - **Technical Context**: Moving averages and RSI provide trend and momentum context - **Advanced Momentum**: MACD, Stochastic, and Williams %R provide sophisticated momentum analysis - **Volatility Measurement**: Bollinger Bands and volatility metrics capture market uncertainty - **Market-Wide Context**: VIX, DXY, and Treasury rates provide macro-economic perspective - **External Factors**: News sentiment incorporates fundamental and event-driven influences - **Regulatory Awareness**: SEC, Fed, and Treasury announcements provide policy-driven signals

2. Computational Efficiency - Limited to 24 numerical features (Phase 1 enhancement from 13) + enhanced textual features from 16+ sources - Avoids redundant indicators that increase noise without adding predictive value - Balances information richness with model complexity while maximizing news coverage and market context - Strategic feature selection focuses on complementary indicators rather than overlapping measurements

3. Data Availability - All features are reliably obtainable from free, public RSS feeds and APIs - Consistent format across different assets (stocks, crypto) with specialized crypto sources - Minimal dependencies on proprietary data feeds with robust fallback mechanisms

---

1.4 TECHNICAL DESIGN DECISIONS

1.4.1 Architecture Rationale

1. Modular Design - **Structured Data Collector**: Isolated technical indicator calculations for easy modification - **Unstructured Data Collector**: Separate news processing pipeline for maintainability - **Main Coordinator**: Centralized orchestration with error handling

2. Data Integration Strategy - **Date Alignment**: Ensures temporal consistency between price and news data - **Quality Assessment**: Validates data completeness and identifies gaps - **Dual Export Format**: CSV for spreadsheet analysis, JSON for programmatic use

1.4.2 Error Handling & Robustness

1. Network Resilience - Enhanced RSS feed processing with 10-second timeouts per source - HTTP retry logic with exponential backoff for API calls - Graceful degradation when individual news sources are unavailable - Comprehensive fallback mechanisms across 11+ RSS feeds

2. Data Validation - Input parameter validation (exchange, symbol format) - Price data completeness checks with technical indicator validation - Enhanced news relevance filtering with duplicate removal - Financial keyword matching to reduce noise and improve signal quality - **Phase 1 Enhancement**: IQR-based outlier detection with configurable thresholds - **Phase 1 Enhancement**: Data completeness scoring and quality metrics - **Phase 1 Enhancement**: Modern pandas methods for missing value handling

---

1.5 VALIDATION RESULTS

1.5.1 Testing Summary

- **AAPL (NYSE)**: Successfully collected complete price data with 24 technical indicators and enhanced news coverage (77 articles)
- **BTC-USD (CRYPTO)**: Confirmed cryptocurrency support with specialized crypto news integration (30 articles)
- **Phase 1 Validation**: Cross-asset testing verified advanced technical indicators, market context, and regulatory news integration
- **Data Quality**: 100% completeness ratio with comprehensive outlier detection and validation
- **News Integration**: 16+ RSS feeds providing diverse coverage including regulatory sources

1.5.2 Data Quality Metrics

- **Structured Data**: 100% completeness for all tested symbols with 24 robust technical indicators
- **Advanced Technical Indicators**: Successfully calculated MACD, Stochastic, Williams %R, and market context indicators
- **Market Context Integration**: Real-time VIX (15.66), DXY (96.928), Treasury 10Y (4.076%) successfully collected
- **News Integration**: Enhanced coverage from 16+ RSS feeds including regulatory sources

- **Phase 1 Validation**: IQR-based outlier detection with zero data quality issues
- **Export Functionality**: Both CSV and JSON formats with comprehensive 24-feature data alignment

1.6 Conclusion

The FinTech Data Curator implements a minimal yet comprehensive feature set that captures the essential elements needed for next-day price prediction:

✅ **Advanced Technical Analysis**: Core OHLCV data enhanced with sophisticated indicators (MACD, Stochastic, Williams %R, RSI, Bollinger Bands, MA)
✅ **Market-Wide Context**: Real-time market indicators (VIX, DXY, Treasury 10Y) providing macro-economic perspective
✅ **Comprehensive News Integration**: 16+ RSS sources including regulatory feeds (SEC, Fed, Treasury) for policy-driven signals
✅ **Enhanced Data Quality**: IQR-based outlier detection, completeness scoring, and robust validation pipeline
✅ **Production-Ready Architecture**: Modular design supporting multiple asset classes with 24-feature structured data
✅ **Regulatory Awareness**: Government and regulatory RSS feeds for early policy impact detection

This enhanced feature set (Phase 1 implementation) strikes an optimal balance between predictive power and computational efficiency, providing a sophisticated foundation for financial machine learning applications. The 24-feature structured dataset combined with 16+ news sources creates a comprehensive data collection system suitable for advanced next-day price prediction models while remaining maintainable and extensible for future enhancements.

References: - Bollinger, J. (2001). Bollinger on Bollinger Bands. McGraw-Hill - Murphy, J. J. (1999). Technical Analysis of the Financial Markets. New York Institute of Finance - Wilder, J. W. (1978). New Concepts in Technical Trading Systems. Trend Research