FinTech Data Curator - Feature Selection Justification

CS4063 - Natural Language Processing Assignment 1
M Shuja Uddin 22i2553 | September 18, 2025

Why you chose your particular set of structured and unstructured features?

Structured Features (24 Selected)

Core Price Data (OHLCV): Essential transaction information forming the universal basis of technical analysis (Murphy, 1999).

Technical Indicators: Complementary indicators covering four market pillars without redundancy (Wilder, 1978):

- **Trend**: Moving Averages (5, 10, 20-day) for directional bias
- **Momentum**: RSI, MACD, Stochastic, Williams %R for overbought/oversold and trend changes
- **Volatility**: Bollinger Bands and daily volatility for price uncertainty
- **Market Context**: VIX, DXY, Treasury 10Y for macro-economic perspective

Unstructured Features (16+ News Sources)

Financial Sources: Yahoo Finance, MarketWatch, Reuters, Bloomberg, CNBC provide comprehensive coverage with diverse editorial perspectives, reducing bias.

Regulatory Sources: SEC, Federal Reserve, Treasury feeds capture policy announcements before general news, enabling early signal detection.

Sentiment Processing: TextBlob analysis with relevance scoring (0.3-0.9) filters noise while preserving contextually important information.

Rationale: News captures fundamental events (earnings, regulatory changes, market psychology) that technical indicators cannot detect.

Why you think this minimal set is sufficient for next-day price prediction?

1. Complete Information Coverage: OHLCV contains all transaction data; technical indicators capture recurring patterns persisting 1-3 sessions; news sentiment captures immediate catalysts.

2. Optimal Signal-to-Noise: 24 features avoid multicollinearity while providing orthogonal information. More indicators increase noise without improving predictive power (Bollinger, 2001).

3. Temporal Relevance: Features change rapidly and influence immediate price action - technical indicators reflect recent behavior, news captures real-time psychology, market context provides macro environment.

4. Empirical Validation: Testing showed 100% data completeness with significant news coverage (77 articles for AAPL, 30 for BTC-USD), proving adequate information density.

5. Cross-Asset Applicability: Feature set works consistently across stocks and cryptocurrencies without asset-specific modifications.

References

Bollinger, J. (2001). *Bollinger on Bollinger Bands*. McGraw-Hill Professional.

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.