**FinTech Data Curation: Features and Justification**

**NLP Assignment – 1 Abdullah Daoud (22I-2626) SE-A**

**Why these features?**

**Structured (market-derived)**

* **OHLCV (Open, High, Low, Close, Volume):** Captures trading range, closing pressure, and liquidity. Basis for indicators and interpretability.
* **Daily Return:** First-order price change; canonical target proxy for short-horizon models.
* **Volatility (20-day rolling std of returns):** Encodes risk regime; volatility clustering aids calibration.
* **SMA\_5 and SMA\_20:** Short- and medium-horizon trend filters summarizing momentum vs. mean-reversion.
* **RSI\_14:** Bounded momentum oscillator to identify exhaustion/overbought-oversold conditions.

**Unstructured (news-derived)**

* **Daily news headlines list** from Yahoo Finance, Google News RSS, and CoinDesk: Captures exogenous signals (earnings, regulation, macro surprises).
* **Lightweight sentiment score (keyword-based):** Transparent, fast polarity proxy resilient to scraping variability.

These choices are intentionally minimal and orthogonal: price level/range (OHLC), participation (Volume), direction (Return), regime (Volatility), trend/momentum (SMA, RSI), and exogenous shocks (Headlines + Sentiment). They avoid redundancy while covering the principal drivers of short-horizon dynamics.

**Why this minimal set is sufficient for next-day prediction?**

* Momentum and trend persistence encoded via SMAs and RSI.
* Volatility clustering stabilizes forecasts and improves calibration.
* Liquidity and attention captured via Volume and Returns.
* Event sensitivity through headlines and sentiment signals.
* Parsimony and robustness: fewer, interpretable features generalize better.
* Operational fit: Features sourced from live web APIs and RSS feeds with retries and normalization.

**References**

* Jegadeesh & Titman (1993). Returns to Buying Winners and Selling Losers.
* Brock, Lakonishok & LeBaron (1992). Simple Technical Trading Rules.
* Wilder (1978). New Concepts in Technical Trading Systems (RSI).
* Engle (1982); Bollerslev (1986). ARCH/GARCH models (Volatility).
* Antweiler & Frank (2004). Is All That Talk Just Noise?
* Financial Innovation (2022). Evidence on combined technical + sentiment features.
* Data: Yahoo Finance API, yfinance, Yahoo/Google News RSS, CoinDesk RSS.