Memoria Técnica – Proyecto Neurotrader

Sistema de Trading Algorítmico Híbrido (Indicadores Técnicos + Sentimiento + ML Tradicional + BiLSTM con Atención)

Autoría: Equipo KeepCoding – Liderado por Darío Fernando Tomatis

Fecha: 08/09/2025

1. Resumen ejecutivo

Este proyecto construye un sistema de trading algorítmico para BTC/USD integrando cuatro capas de señal: (1) indicadores técnicos clásicos, (2) modelos de Machine Learning tradicionales (Random Forest, SGD y XGBoost), (3) análisis de sentimiento de noticias/redes, y (4) un modelo secuencial *BiLSTM con atención* que actúa como integrador final de señales. Se trabajó con datos diarios (2020–2025), backtesting y validaciones *rolling out-of-sample*, calibración de probabilidades y un pipeline reproducible para análisis y dashboard.

2. Objetivos

- Desarrollar un sistema híbrido que combine múltiples fuentes de información para generar señales de compra/venta.
- Asegurar una evaluación rigurosa mediante splits temporales, calibración de probabilidades y backtesting.
- Entregar artefactos reutilizables (datasets enriquecidos, scripts de entrenamiento y backtesting, KPIs y reportes).

3. Datos y fuentes

3.1 Mercado

- BTC/USD diario (OHLCV): 2020–2025.
- Archivo base enriquecido: `btc_enriched_indicators.csv` / `btc_enriched_with_target.csv`.
- Columnas principales: Date, Open, High, Low, Close, Volume + indicadores: RSI, MACD, MACD_SIGNAL, SMA20, EMA20, BB_UPPER, BB_LOWER, ATR, CCI.
- Variable objetivo (`y_true`): 1 si el cierre de mañana > cierre de hoy, 0 en caso contrario.

3.2 Sentimiento

- Noticias/tweets de BTC procesados con un modelo tipo FinBERT (o similar) para obtener: `proba_sentiment_neg`, `proba_sentiment_neu`, `proba_sentiment_pos`.
- Almacenados/mergeados en: `btc_merged_with_sentiment_clean.csv`.

4. Ingeniería de características

- Indicadores técnicos calculados con *finta* y verificación visual de consistencia (precio vs medias, RSI, MACD).
- Probabilidades calibradas de modelos base incorporadas como *features* para el integrador secuencial (BiLSTM con atención).
- Features de sentimiento (neg, neu, pos) como insumos complementarios.

5. Modelos y roles

5.1 Random Forest Classifier (RFC)

- Ensemble de árboles, robusto a no linealidades y ruido.
- Entrenamiento con indicadores técnicos. *class_weight* para posible desbalance.
- Calibración de probabilidades vía regresión isotónica.
- Esquema *rolling* mensual (2021–2024) y actualizaciones periódicas en 2025.

5.2 SGDClassifier (SGD)

- Modelo lineal con *Stochastic Gradient Descent*, rápido y actualizable.
- *Pipeline* con estandarización; calibración con Platt (sigmoid).
- Uso de *partial_fit* para aprendizaje incremental diario.

5.3 XGBoost (como generador de proba feature)

- Gradient boosting en árboles para producir `proba_xgb_cal` incorporada al integrador secuencial.
- Calibración posterior para interpretabilidad probabilística.

5.4 Modelo de Sentimiento (FinBERT o equivalente)

- Transformers *fine-tuned* para dominio financiero.
- Salidas: `proba_sentiment_neg`, `proba_sentiment_neu`, `proba_sentiment_pos`.

5.5 Integrador secuencial: BiLSTM con atención

- Arquitectura: BiLSTM (64 → 32 por dirección) + mecanismo de atención + capa densa final.

- Entrada: ventanas fijas de 5 días con indicadores + proba calibradas de RFC/SGD/XGB + sentimiento.
- Salida: probabilidad de subida (0–1). Reglas de señal (ejemplo): comprar si >= 0.60, salir si < 0.45 (ajustable).
- Embargo operativo: no se predice el día 1 de cada mes (ventanas inconsútiles).

6. Metodología de validación

- **Split temporal**: Entrenamiento base 2015–2020; Validación 2021-H1; Test 2021-H2 → Selección + calibración de probas.
- **Rolling OOS 2021–2024**: generación de probabilidades out-of-sample por mes y consolidación en CSV.
- **Backtesting**: estrategia por umbrales, capital inicial 1.000 USD, posición fraccional y costes de transacción.

6.1 Resultados resumidos (histórico RFC/SGD)

- Backtest simple (2021–2024): - RFC: equity final ~1.036 ×, 34 operaciones, win rate ~53%. - SGD: equity final ~1.068 ×, 10 operaciones, win rate ~40%.
- Optimización de umbrales/fracción de capital: Mejor combinación: SGD con *entry_thr*=0.6, *exit_thr*=0.4, *pos_frac*=0.2 → equity ~3.185 ×.

6.2 Resultados 2025 (pipeline LSTM)

Periodo evaluado: 2025-01-05 a 2025-08-30 (210 días). Métricas reportadas en una simulación:

- Operaciones: 4
- % de días en mercado: ~11.4%
- Equity final estrategia: 0.8723 ×
- Equity buy&hold: 1.0988 ×
- Máximo *drawdown*: -12.77%
- Sharpe (252): -2.15

Interpretación breve: el desempeño 2025 fue inferior a *buy&hold* en este corte; consistente con fases de subajuste/umbrales no óptimos y bajo número de operaciones. Muestra la necesidad de sintonía fina por régimen y de stress-tests adicionales.

7. Pipeline operativo (día a día)

1. Descarga/actualización de precios y noticias del día.
2. Cálculo de indicadores técnicos y merge de sentimiento.
3. Obtención de proba calibradas (RFC/SGD/XGB) y armado de secuencias (5 días).
4. Inferencia con BiLSTM con atención → probabilidad de compra.

5. Conversión a señal con reglas de umbral y gestión de capital; registro para backtesting y dashboard.
6. Actualización incremental del SGD con *partial_fit*.

8. Artefactos y salidas

- CSVs estandarizados (para backtesting y dashboard): columnas con OHLCV, indicadores, proba calibradas, sentimiento y `y_true`.
- KPIs de backtesting anuales y globales; archivos de resultados por escenario.
- Notebooks y scripts de entrenamiento/inferencia con configuración reproducible.

9. Reproducibilidad

- Estructura de repositorio: módulos separados para datos, features, modelos base, integrador LSTM, backtesting y dashboard.
- Convenciones de archivo y nombres coherentes (ej.: `btc_2025_with_probs.csv`).
- Uso de ficheros "placeholder" (`.gitkeep`) para versionar carpetas vacías.

10. Buenas prácticas y riesgos

- Evitar *data leakage*: calibración/optimización sólo con datos previos a test.
- Control de *look-ahead bias* y artefactos de ventana (embargos y bordes de mes).
- Calibración sistemática de probabilidades antes de agregarlas al LSTM.
- Costes de transacción y deslizamiento contemplados en escenarios.

11. Conclusiones

- La arquitectura híbrida demuestra viabilidad: los modelos clásicos aportan señal base robusta y el BiLSTM con atención integra contexto temporal y de sentimiento.
- La calibración de probabilidades y el backtesting *rolling* acercan el pipeline a un estándar profesional.
- Los resultados favorables con SGD tras optimización sugieren valor en estrategias de baja complejidad bien calibradas.
- Los cortes de 2025 evidencian sensibilidad a umbrales/régimen: se recomienda sintonía por régimen y validaciones intradía.

12. Roadmap de mejoras

1. Datos intradía (1h/4h/15m) y *features* multi-horizonte.
2. Optimización y *tuning* por régimen de mercado (volatilidad, tendencia, halving windows).

3. Revisión de reglas de entrada/salida y *position sizing* con control de riesgo (ATR stops, VaR, Kelly fraccional).
4. Automatización de despliegue e inferencia en tiempo real (bridge/API a MetaTrader/Exchange).
5. Seguimiento con MLflow, tests unitarios del backtester y validaciones cruzadas temporales.

13. Equipo y organización

Equipo de 5 integrantes con roles rotativos: adquisición de datos, ingeniería de características, modelos base (RFC/SGD/XGB), NLP de sentimiento, integrador BiLSTM con atención, backtesting y dashboard; coordinación con enfoque ágil.

14. Referencias y anexos

- Dataset y notebooks provistos durante el desarrollo (KeepCoding) y archivos del proyecto.
- Paper de referencia: "A Comprehensive Analysis of Machine Learning Models for Algorithmic Trading of Bitcoin", arXiv:2407.18334.
- Anexo A – Lista de columnas estandarizadas y definición de `y_true`.
- Anexo B – Principales CSV de salida para backtesting/dashboard.

Anexo A – Esquema de columnas

Columna	Descripción
Date	Fecha (diaria, zona UTC)
Open, High, Low, Close, Volume	OHLCV diario
RSI, MACD, MACD_SIGNAL	Indicadores de momentum
SMA20, EMA20	Medias móviles
BB_UPPER, BB_LOWER	Bandas de Bollinger
ATR, CCI	Volatilidad / ciclo (composite)

proba_rfc_cal, proba_sgd_cal, proba_xgb_cal	Probabilidades calibradas de modelos base
proba_sentiment_neg/neu/pos	Sentimiento de noticias/tweets
y_true	1 si Close(t+1)>Close(t), 0 en caso contrario

Anexo B – Artefactos de salida

- CSVs de backtesting por año y globales (KPIs y señales).
- CSVs para dashboard (Power BI): métricas y series listas para visualización.