

Análisis Exploratorio, Detección de Anomalías y Modelado LSTM

Jose Ortiz Pablo Sanchez Dayana Henao David Alava Sebastian Agudelo Luis Vera
Maicol Taborda

Análisis Cuantitativo de Mercados

Diciembre 6, 2025

Agenda

- 1 Introducción y Objetivos
- 2 Carga y Preparación de Datos
- 3 EDA: Análisis de Precios del Oro
- 4 EDA: Análisis del Corpus de Noticias
- 5 Detección de Anomalías
- 6 Análisis de Sentimientos con FinBERT
- 7 Correlación y Causalidad de Granger
- 8 Modelado Predictivo con LSTM
- 9 Síntesis y Resultados Finales
- 10 Apéndices

Contexto del Proyecto

Motivación:

- El oro es un activo refugio en tiempos de incertidumbre
- Las noticias financieras influyen en la percepción del mercado
- ¿Puede el sentimiento de noticias predecir movimientos de precio?

Hipótesis Principal:

El sentimiento agregado de noticias del Wall Street Journal tiene poder predictivo sobre los precios del oro

Datos Utilizados

- **Precios:** 3,614 días
- **Noticias:** 18,776 artículos
- **Período:** 2016-2025
- **Precios:** Dukascopy
- **Noticias:** WSJ + BBC

Objetivos del Análisis

① Análisis Exploratorio de Datos (EDA)

- Caracterización estadística de precios del oro
- Análisis del corpus de noticias del WSJ + BBC

② Detección de Anomalías

- Identificación de eventos extremos en precios
- Análisis de contexto en noticias

③ Análisis de Sentimientos

- Clasificación con FinBERT
- Agregación temporal de sentimientos

④ Correlación y Causalidad

- Test de Granger
- Análisis de lags óptimos

⑤ Modelado Predictivo

- LSTM con y sin sentimiento
- Evaluación comparativa

Notebook 01: Introducción y Carga de Datos

Precios del Oro (OHLCV):

- Fuente: Dukascopy Bank SA (XAU/USD)
- Frecuencia: Diaria
- Variables: Open, High, Low, Close, Volume
- Período: 2016-01-03 a 2025-11-24
- Total: **3,614 observaciones**

Procesamiento:

- Conversión de fechas a datetime
- Detección de valores faltantes (0 %)
- Cálculo de retornos logarítmicos
- Cálculo de volatilidad histórica

Noticias WSJ + BBC:

- Fuente: Web scraping (Wall Street Journal + BBC)
- Variables: Título, fecha, hipervínculo
- Período: 2012-2025
- Total: **18,776 artículos**

Limpieza:

- Eliminación de duplicados
- Normalización de texto
- Filtrado por palabras clave (gold, metal)
- Parsing de fechas

Resultado

Dataset integrado final: **2,273 días** con precios y noticias asociadas

Notebook 02: EDA de Precios del Oro

Estadísticas Descriptivas:

Métrica	Valor	Unidad
Media	1,806.03	USD
Mediana	1,753.49	USD
Desv. Estándar	621.50	USD
CV	34.4 %	—
Mínimo	1,063.06	USD
Máximo	4,365.22	USD
Rango	3,302.16	USD
Skewness	+0.65	—
Kurtosis	-0.52	—

Interpretación:

- Alta volatilidad ($CV = 34.4 \%$)
- Distribución sesgada a la derecha
- Colas más ligeras que normal



Figura: Serie temporal de precios del oro (2016-2025)

Tendencias observadas:

- Crecimiento sostenido 2019-2020
- Pico histórico en 2020 (COVID-19)

Análisis de Distribuciones

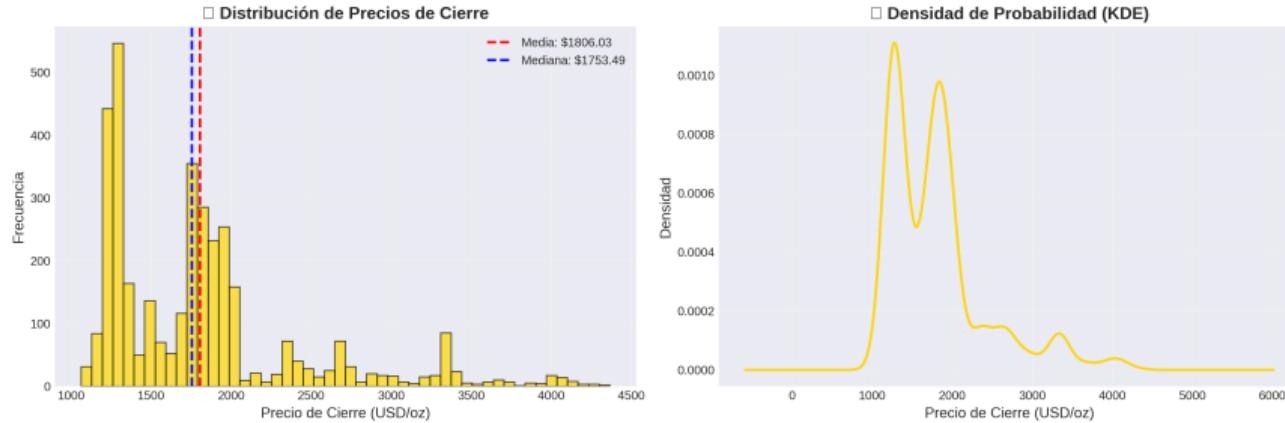


Figura: Distribución de precios de cierre del oro

Test de Normalidad (Jarque-Bera):

- $p\text{-value} < 0,001 \rightarrow$ Rechazamos normalidad
- Distribución no gaussiana con asimetría positiva

Análisis de Retornos y Tests Estadísticos

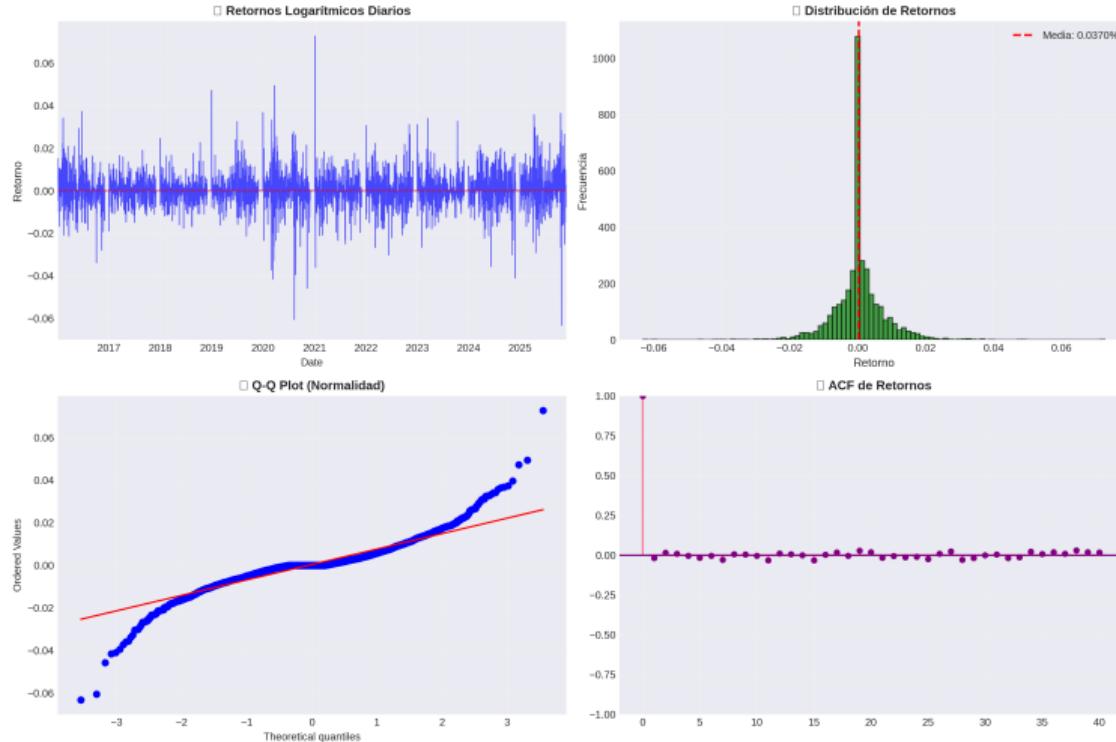


Figura: Retornos logarítmicos del oro

Descomposición de Series Temporales

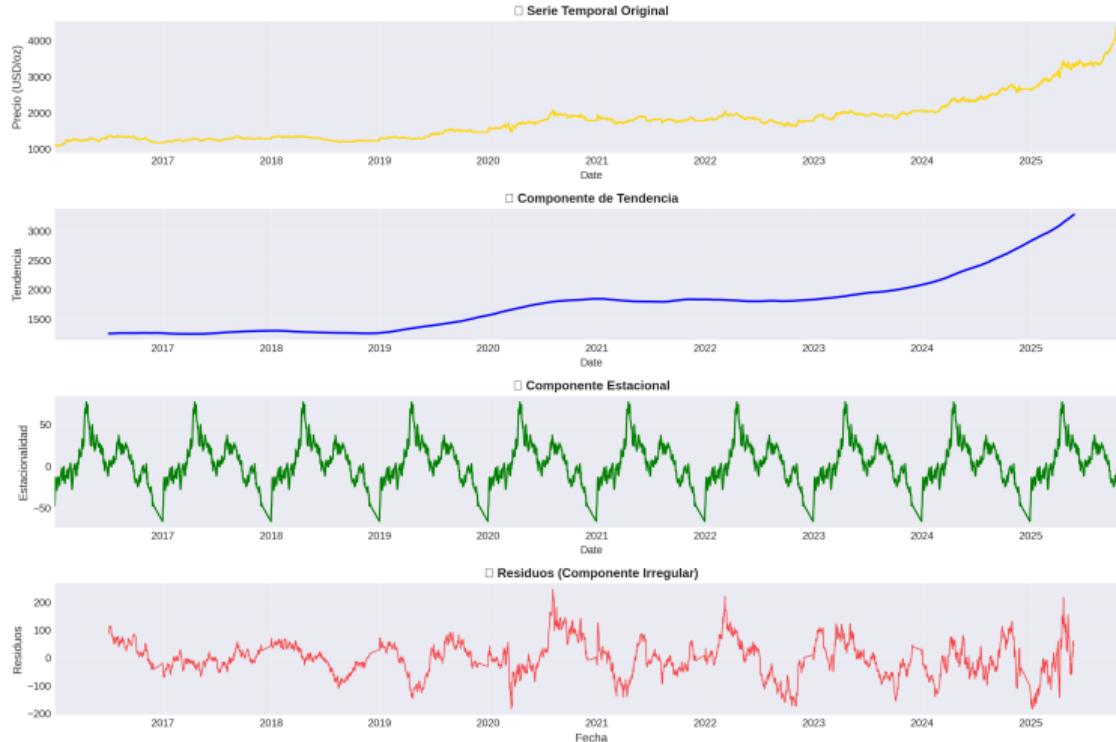


Figura: Descomposición STL: Tendencia + Estacionalidad + Residuales

Correlación OHLCV

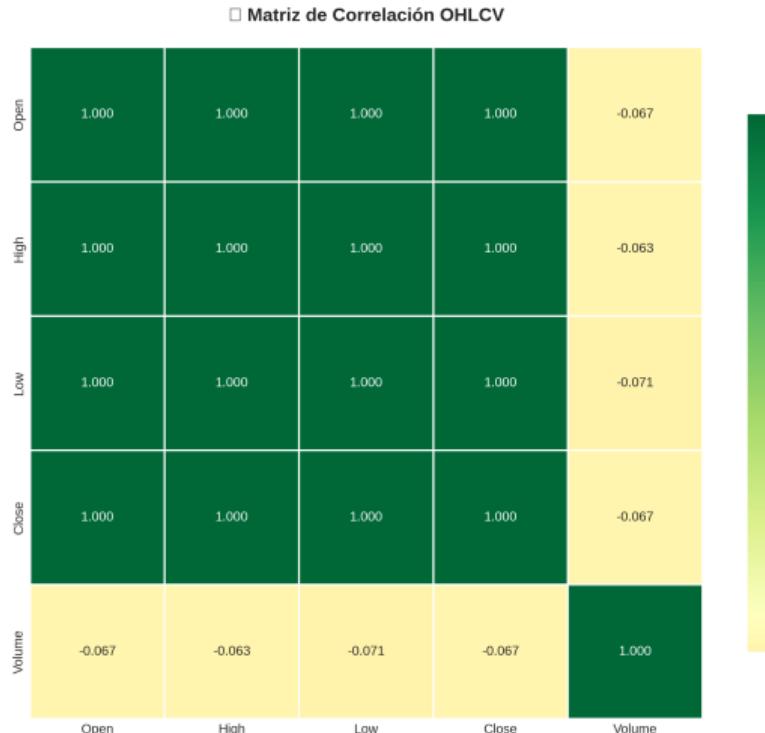


Figura: Matriz de correlación de variables OHLCV

Notebook 03: EDA de Noticias WSJ + BBC

Estadísticas del Corpus:

- Total artículos: 18,776
- Período: 2012-2025 (13 años)
- Promedio diario: 3.5 noticias
- Artículos únicos: 18,532
- Tasa duplicación: 1.3 %

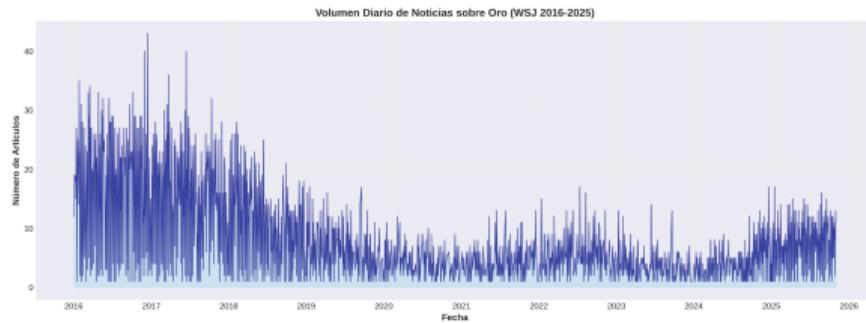


Figura: Volumen diario de noticias sobre oro

Procesamiento NLP:

- Tokenización con spaCy
- Eliminación de stopwords
- Lemmatización
- Extracción de entidades (NER)

Patrones temporales:

- Picos en crisis (COVID-19, 2020)
- Mayor cobertura 2022-2024
- Media móvil estable

Análisis de Títulos

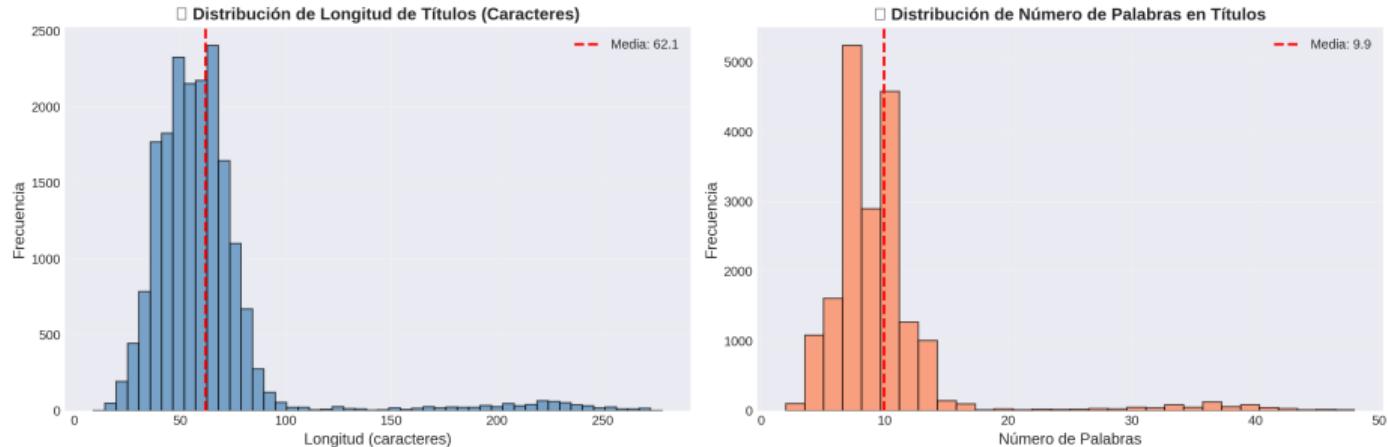


Figura: Distribución de longitud de títulos

Estadísticas: Media 12.5 palabras — Mediana 11 palabras — Rango 3-35 palabras

Análisis de Palabras Frecuentes

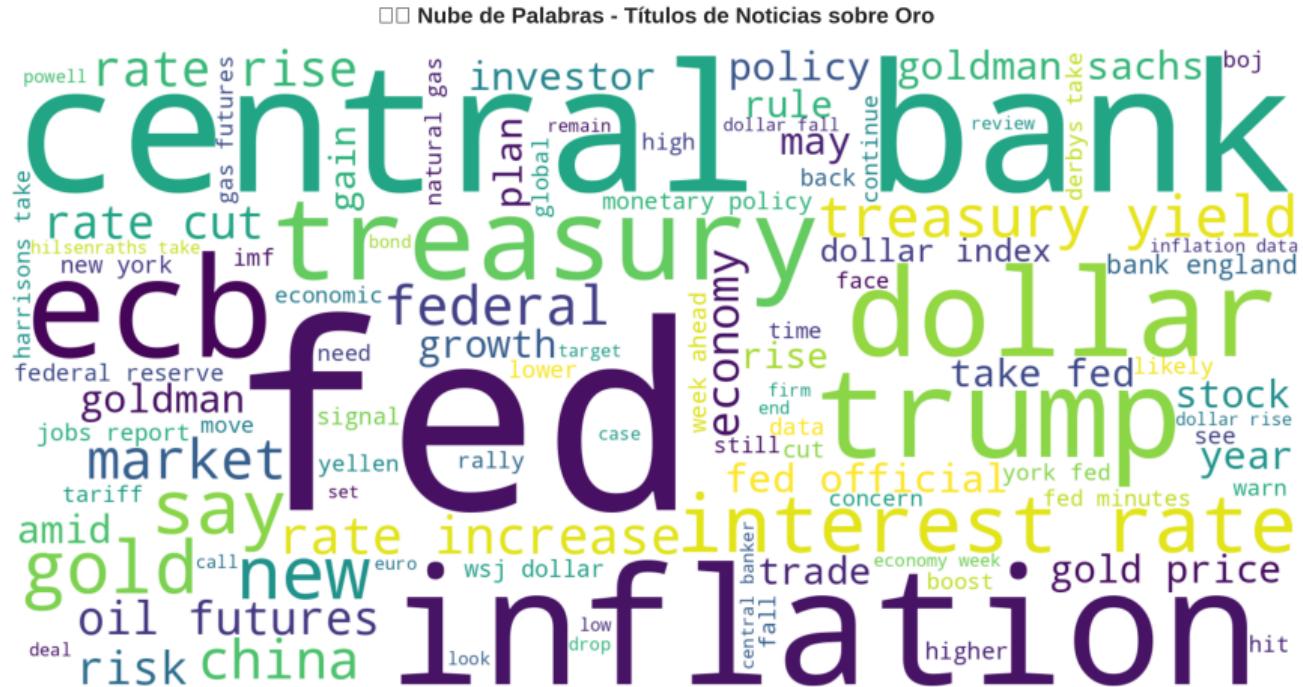


Figura: Nube de palabras más frecuentes en títulos

Términos dominantes: gold, price, market, rise, fall, trading, investors, demand, dollar, inflation

Notebook 04: Detección de Anomalías en Precios

Métodos: IQR, Z-Score, Isolation Forest

Criterio: Consenso de al menos 2 de 3 métodos

Resultados:

Método	Outliers	%
IQR	258	7.1
Z-Score	69	1.9
Isolation Forest	181	5.0
Consenso	132	3.7

Principales eventos: COVID-19 (Mar 2020), Máximo histórico (Ago 2020), Invasión Ucrania (Feb 2022), Nuevo máximo (Mar 2024)

Modelo: FinBERT (ProsusAI/finbert)

- Clasificación: Positivo, Neutral, Negativo
- Input: Títulos de noticias WSJ + BBC
- Agregación diaria con medias móviles (7 y 30 días)

Distribución de Sentimientos

Clase	Proporción
Positivo	21.8 %
Neutral	42.6 %
Negativo	35.6 %

Estadísticas: Media -0.127 — Desviación 0.284 — Rango [-0.970, +0.948]

Notebook 06: Análisis de Correlación

Correlación Cross-Lag: Sentimiento vs Retornos en lags $k \in [-30, +30]$

Lag (días)	Correlación	p-value
-6	-0.070	0.001
0	-0.015	0.485
+6	+0.024	0.278

Resultado: Lag óptimo = -6 días ($\rho = -0,070$) — Correlación débil

Test de Causalidad de Granger

Hipótesis: H_0 : Sentimiento NO mejora predicción de retornos

Dirección	Lags	p-value
Sentiment → Returns	5	0.184
Sentiment → Returns	10	0.267

Conclusión

NO se encontró causalidad de Granger significativa ($p > 0.05$)

Notebook 07: Modelo LSTM Integrado

Arquitectura:

- 3 capas LSTM (256-128-64 units) + Dropout (0.2)
- Optimizador: Adam ($lr=0.001$) — Loss: MSE
- Sequence length: 60 días
- División: 60 % Train / 20 % Val / 20 % Test

Features:

- Base: OHLC + Returns + Volatility
- Sentiment: Base + Sentiment + MA7 + MA30

Comparación de Modelos LSTM

Modelo	RMSE	MAE	R ²	MAPE
LSTM Base	472.94	406.75	-0.39	12.65 %
LSTM + Sentiment	634.65	563.84	-1.50	17.71 %
Diferencia	-34.2 %	-38.6 %	-285 %	-40.0 %

Conclusión

Sentimiento **empeora** desempeño predictivo

Notebook 08: Síntesis de Resultados

Resultados Principales:

- **Datos:** 3,614 días precios — 18,776 noticias — 132 anomalías
- **EDA:** CV = 34.4 % (alta volatilidad) — 42.6 % noticias neutrales
- **Correlación:** $\rho = -0,070$ (débil)
- **Causalidad:** Test Granger NO significativo ($p > 0.05$)
- **LSTM:** Sentimiento empeora predicciones (-34 % RMSE)

Conclusiones Principales

Hipótesis Principal: RECHAZADA

Sentimiento de noticias WSJ + BBC **NO** tiene poder predictivo significativo

Hallazgos:

- Correlación débil e inconsistente
- Sin causalidad de Granger
- Sentimiento empeora LSTM
- Anomalías correlacionan con eventos globales

Posibles razones:

- Mercado eficiente (información ya incorporada)
- Factores macro dominan sobre sentimiento
- Fuente única insuficiente (WSJ + BBC no es suficiente)

Limitaciones del Estudio

- ① **Datos:** Solo títulos — Fuente única (WSJ + BBC)
- ② **Modelo:** FinBERT genérico — Agregación diaria
- ③ **Variables:** Sin indicadores macro ni redes sociales
- ④ **Arquitectura:** Solo LSTM — Sin Transformers/GRU

- ① **Datos:** Múltiples fuentes (Reuters, Bloomberg) — Contenido completo — Redes sociales
- ② **Métodos:** Modelos específicos commodities — Granularidad intraday
- ③ **Features:** Indicadores macro (tasas, inflación, DXY)
- ④ **Modelos:** Transformers — Ensemble learning

Apéndice A: Stack Tecnológico

Core:

- Python 3.9+
- pandas, numpy, scipy
- scikit-learn, statsmodels

Visualización:

- matplotlib, seaborn, plotly

NLP/DL:

- transformers (HuggingFace)
- PyTorch, Keras
- spaCy

Datos:

- yfinance
- beautifulsoup4

¡Gracias por su atención!

Proyecto: Predicción de Precios del Oro con Análisis de Sentimientos

Repositorio: <https://github.com/dassjoss/PredicticModel/>

Diciembre 6, 2025