

Efecto del Sentimiento en Redes Sociales sobre el Precio del Bitcoin: Un Análisis Comparativo de Modelos de NLP y Técnicas de Reducción de Dimensionalidad

Sebastián Marinovic Leiva¹ Ricardo Iván Lizana Aseña¹
Luis Andrés Gutiérrez González¹

¹Magíster en Data Science, Universidad de Las Américas, Chile

Enero 2026

Resumen

El presente estudio investiga la capacidad predictiva del sentimiento expresado en Twitter sobre el precio de Bitcoin, comparando sistemáticamente métodos tradicionales de análisis de sentimiento (VADER, TextBlob) con modelos de procesamiento de lenguaje natural basados en Transformers (FinBERT, Twitter-RoBERTa). Adicionalmente, se evalúa el impacto de la reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) sobre las variables de sentimiento. Utilizando un corpus de 904,427 tweets y datos de precios horarios durante el período febrero-agosto 2021, se entrenaron modelos de regresión (Ridge, Random Forest) y redes neuronales LSTM bajo seis escenarios de features. Los resultados revelan que: (1) FinBERT logra una correlación significativamente más fuerte con los retornos ($\rho = 0,113$, $p < 0,001$) en comparación con VADER ($\rho = 0,027$, no significativo); (2) la combinación PCA + FinBERT con LSTM alcanza el mejor desempeño predictivo (MAE = 0.0316), mejorando un 18.6 % respecto al modelo base; (3) el PCA combinado de múltiples métodos (PCA_ALL) obtiene el mejor resultado para Random Forest (MAE = 0.0586). Estos hallazgos sugieren que tanto la elección del método de extracción de sentimiento como la técnica de reducción de dimensionalidad impactan significativamente en la capacidad predictiva de los modelos.

Palabras clave: Bitcoin, análisis de sentimiento, FinBERT, PCA, LSTM, predicción de precios, Twitter, NLP, machine learning.

1. Introducción

Bitcoin (BTC) representa el activo digital más relevante del ecosistema de criptomonedas, caracterizado por una volatilidad extrema y sensibilidad a factores exógenos [1]. A diferencia de los mercados financieros tradicionales, el precio de Bitcoin puede verse influenciado significativamente por el sentimiento colectivo expresado en redes sociales, donde publicaciones virales pueden coincidir con movimientos de precio sustanciales [2].

El análisis de sentimiento ha emergido como una herramienta fundamental para cuantificar las percepciones del mercado a partir de texto no estructurado [3, 4]. Tradicionalmente, métodos basados en léxicos como VADER [5] han sido ampliamente utilizados debido a su simplicidad y velocidad de ejecución. Sin embargo, estos enfoques presentan limitaciones al no capturar el contexto específico del dominio financiero, donde términos como “bearish”, “bullish” o “HODL” poseen significados particulares.

La aparición de modelos de lenguaje basados en arquitecturas Transformer [7] ha revolucionado el campo del NLP. FinBERT [8], un modelo pre-entrenado específicamente con textos

financieros, promete una mejor comprensión del lenguaje especializado de los mercados. Paralelamente, modelos como Twitter-RoBERTa [9] han sido optimizados para el estilo particular de las redes sociales.

Un desafío adicional en el modelado predictivo con variables de sentimiento es la alta dimensionalidad y potencial multicolinealidad entre diferentes métricas de sentimiento. El Análisis de Componentes Principales (PCA) ofrece una solución para reducir la dimensionalidad preservando la información más relevante, potencialmente mejorando la generalización de los modelos [11].

Este trabajo presenta las siguientes contribuciones:

1. Comparación sistemática de cuatro métodos de análisis de sentimiento: VADER, TextBlob, FinBERT y Twitter-RoBERTa.
2. Evaluación del impacto de PCA aplicado a variables de sentimiento bajo múltiples configuraciones.
3. Análisis de seis escenarios de features con tres familias de modelos predictivos.
4. Identificación de la combinación óptima (LSTM + PCA_FinBERT) que logra una mejora del 18.6 % sobre el baseline.

2. Marco Teórico

2.1. Volatilidad de Bitcoin y señales sociales

Bitcoin exhibe una volatilidad significativamente mayor que activos tradicionales, con variaciones diarias que frecuentemente superan el 5 %. Kristoufek [1] documentó una relación significativa entre búsquedas en Google Trends y Wikipedia con la dinámica de precios de BTC. Garcia y Schweitzer [2] identificaron patrones de retroalimentación entre indicadores sociales y movimientos del mercado.

2.2. Análisis de sentimiento basado en léxicos

VADER (Valence Aware Dictionary and sEntiment Reasoner) [5] es una herramienta basada en léxico específicamente diseñada para redes sociales. Produce un score compuesto en el rango $[-1, 1]$ que captura la valencia emocional del texto, considerando factores como mayúsculas, puntuación y emojis. TextBlob ofrece funcionalidad similar con métricas adicionales de subjetividad.

2.3. Modelos Transformer para NLP financiero

La arquitectura Transformer [6] y su implementación BERT [7] han establecido nuevos estándares en tareas de NLP. FinBERT [8] adapta BERT mediante fine-tuning con el corpus Financial PhraseBank y textos de Reuters TRC2, logrando capturar matices específicos del lenguaje financiero. El modelo produce probabilidades para tres clases (positivo, negativo, neutral), permitiendo calcular un score de sentimiento como:

$$\text{score}_{\text{FinBERT}} = P(\text{positivo}) - P(\text{negativo}) \quad (1)$$

Twitter-RoBERTa [9] es una variante de RoBERTa [10] entrenada con 58 millones de tweets, optimizada para el estilo informal y abreviado característico de las redes sociales.

2.4. Análisis de Componentes Principales

PCA [11] es una técnica de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un conjunto de variables linealmente independientes llamadas componentes principales. Matemáticamente, PCA busca la proyección lineal que maximiza la varianza:

$$\mathbf{z} = \mathbf{W}^T \mathbf{x} \quad (2)$$

donde \mathbf{W} contiene los eigenvectores de la matriz de covarianza ordenados por eigenvalor descendente. Los primeros k componentes capturan la mayor proporción de varianza con mínima pérdida de información.

2.5. Redes LSTM para series temporales

Las redes Long Short-Term Memory (LSTM) [12] son una arquitectura de redes neuronales recurrentes capaces de aprender dependencias de largo plazo. Su estructura de compuertas (input, forget, output) permite mantener información relevante a través de múltiples pasos temporales, haciéndolas particularmente adecuadas para series financieras.

3. Metodología

3.1. Datos

El estudio utiliza dos fuentes de datos principales:

- **Tweets:** 1,000,025 publicaciones relacionadas con Bitcoin obtenidas de Kaggle, de las cuales 904,427 pasaron los filtros de calidad tras eliminar duplicados y textos demasiado cortos.
- **Precios:** Serie OHLCV (Open, High, Low, Close, Volume) con frecuencia horaria de Binance para el par BTC/USDT.

Cuadro 1: Estadísticas descriptivas del conjunto de datos

Variable	Valor
Período analizado	07-Feb-2021 a 20-Ago-2021
Total tweets (raw)	1,000,025
Tweets tras limpieza	904,427
Horas con tweets	1,523
Observaciones para modelado	1,449
Promedio tweets/hora	594

3.2. Extracción de sentimiento

Se aplicaron cuatro métodos de análisis de sentimiento a nivel de tweet individual:

1. **VADER:** Score compound $[-1, 1]$ más componentes neg, neu, pos.
2. **TextBlob:** Polaridad $[-1, 1]$ y subjetividad $[0, 1]$.
3. **FinBERT:** Modelo ProsusAI/finbert con score calculado según Ec. (1).
4. **Twitter-RoBERTa:** Modelo cardiffnlp/twitter-roberta-base-sentiment-latest.

Adicionalmente, se extrajeron indicadores FOMO (Fear Of Missing Out) y FUD (Fear, Uncertainty, Doubt) mediante conteo de palabras clave específicas del dominio crypto.

3.3. Agregación horaria y features

Los scores de sentimiento a nivel de tweet fueron agregados a nivel horario mediante estadísticas descriptivas (media, desviación estándar, mínimo, máximo). Se generaron features adicionales incluyendo:

- **Features técnicos:** Retornos logarítmicos, volatilidad rolling, RSI, momentum.
- **Features de sentimiento:** EWMA del sentimiento, momentum del sentimiento, z-score, divergencia FinBERT-VADER.
- **Features de volumen social:** Conteo de tweets, ratio de volumen de tweets.

3.4. Análisis de Componentes Principales

Se aplicó PCA a tres conjuntos de variables de sentimiento:

1. **PCA_VADER:** 7 variables basadas en VADER y TextBlob.
2. **PCA_FinBERT:** 8 variables incluyendo FinBERT y features derivados.
3. **PCA_ALL:** 11 variables combinando todos los métodos de sentimiento.

Para cada configuración se retuvieron 3 componentes principales, explicando entre 77 % y 84 % de la varianza total.

3.5. Escenarios de modelado

Se definieron seis escenarios de features para evaluación sistemática:

Cuadro 2: Escenarios de features evaluados

Escenario	Descripción	Features
BASE	Solo features técnicos	5
VADER	Técnicos + sentimiento VADER	12
FinBERT	Técnicos + sentimiento FinBERT	15
PCA_VADER	Técnicos + PCA de VADER	8
PCA_FinBERT	Técnicos + PCA de FinBERT	8
PCA_ALL	Técnicos + PCA combinado	8

3.6. Modelos predictivos

Se evaluaron tres familias de modelos:

- **Naive (Zero Return):** Baseline que predice retorno cero.
- **Ridge Regression:** Modelo lineal con regularización L2 ($\alpha = 1,0$).
- **Random Forest:** 200 árboles, profundidad máxima 10.
- **LSTM:** 2 capas, 64 unidades hidden, dropout 0.2, secuencia de 24 horas.

3.7. Evaluación

Se utilizó validación walk-forward con 5 splits para modelos tradicionales, preservando la estructura temporal de los datos. La métrica principal es el Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

Para LSTM se utilizó split temporal 80/20 con early stopping basado en validation loss.

4. Resultados

4.1. Comparación de métodos de sentimiento

La Tabla 3 presenta las estadísticas descriptivas de los métodos de sentimiento aplicados.

Cuadro 3: Comparación de métodos de sentimiento

Método	Media	Std	Observación
VADER	0.245	0.421	Sesgo positivo pronunciado
TextBlob	0.120	0.230	Distribución más centrada
FinBERT	0.061	0.244	Más neutral, apropiado para finanzas
RoBERTa	0.221	0.437	Similar a VADER

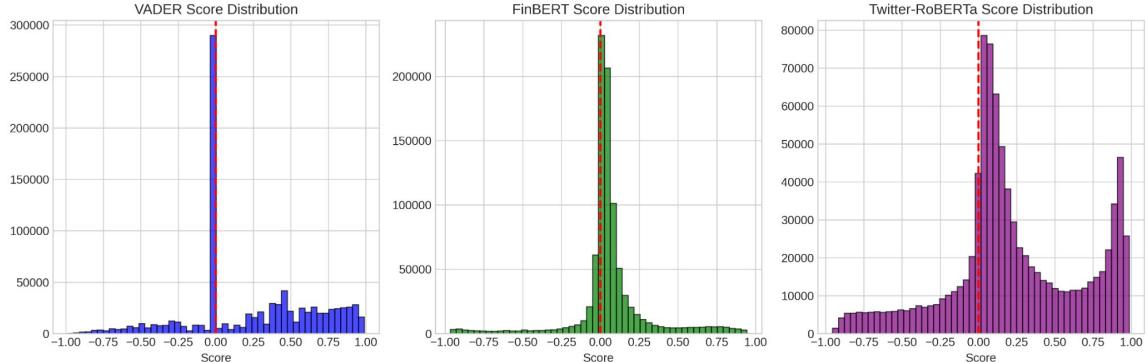


Figura 1: Distribución del sentimiento según VADER, FinBERT y Twitter-RoBERTa. FinBERT muestra una distribución más centrada y neutral, consistente con expectativas de mercados eficientes.

La correlación entre métodos (Figura 2) revela que FinBERT captura información parcialmente diferente a los métodos basados en léxico, con correlaciones moderadas (VADER-FinBERT: $r = 0,409$).

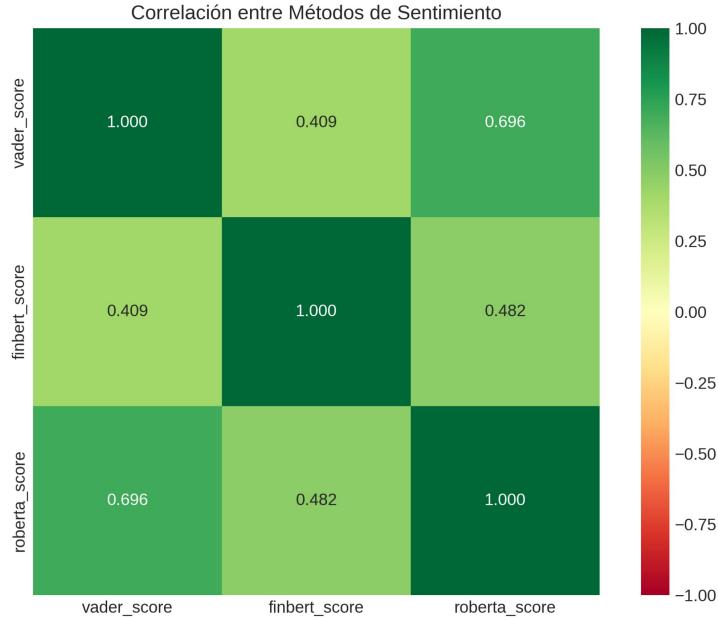


Figura 2: Matriz de correlación entre métodos de sentimiento.

4.2. Correlación sentimiento-retornos

La Tabla 4 presenta el hallazgo más relevante: FinBERT logra una correlación significativamente más fuerte y estadísticamente significativa con los retornos de Bitcoin.

Cuadro 4: Correlación Spearman entre sentimiento y retornos (24h)

Método	ρ	p-value	Significancia
VADER	0.027	0.302	No significativo
FinBERT	0.113	1.53e-05	Muy significativo (***)
PCA_ALL PC1	0.084	1.47e-03	Significativo (**)
PCA_ALL PC3	-0.098	1.77e-04	Muy significativo (***)

FinBERT logra una correlación **4.2 veces más fuerte** que VADER. Notablemente, el tercer componente principal del PCA combinado muestra correlación negativa significativa, sugiriendo que captura un factor de sentimiento inversamente relacionado con los retornos.

4.3. Análisis de Componentes Principales

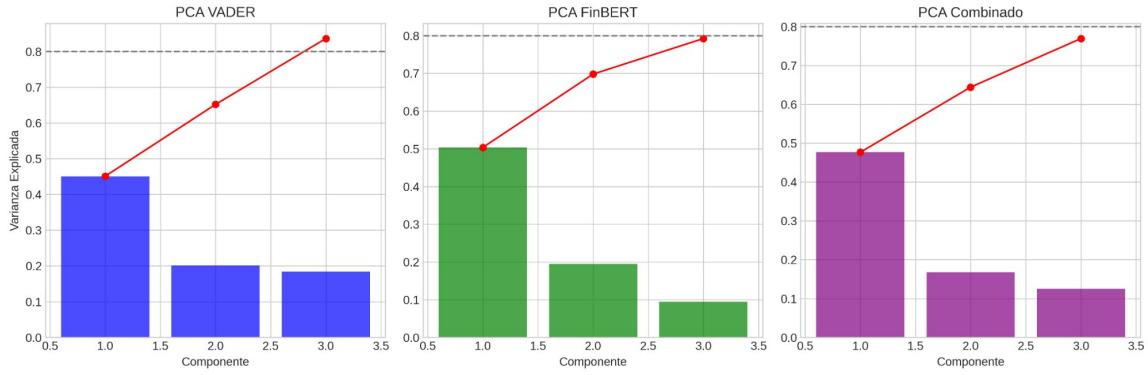


Figura 3: Varianza explicada por los componentes principales en cada configuración de PCA.

Cuadro 5: Varianza explicada acumulada por configuración de PCA

Configuración	PC1	PC1-PC2	PC1-PC3
PCA_VADER	45.1 %	65.2 %	83.6 %
PCA_FinBERT	50.4 %	69.8 %	79.2 %
PCA_ALL	47.7 %	64.5 %	76.9 %

La Figura 4 muestra los loadings del PCA combinado, revelando la contribución de cada variable a los componentes principales.

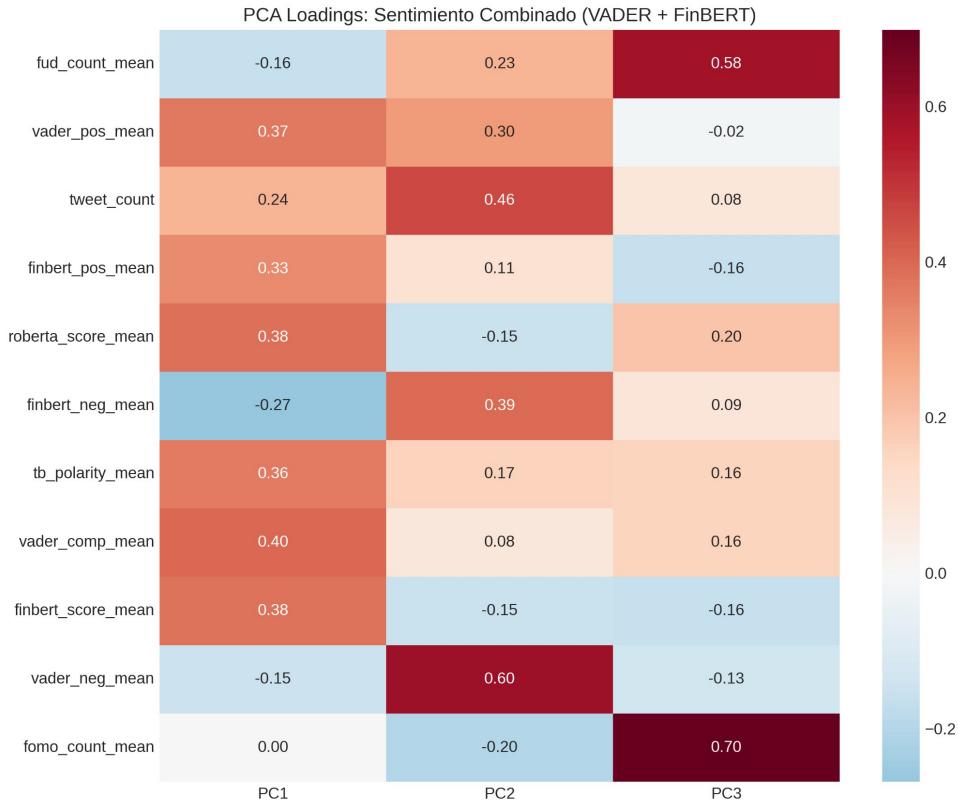


Figura 4: Loadings del PCA combinado (VADER + FinBERT + RoBERTa). PC1 captura el sentimiento general positivo, PC2 el volumen y sentimiento negativo, PC3 las señales FOMO/FUD.

La interpretación de los componentes principales:

- **PC1 (47.7 % varianza):** “Índice de sentimiento positivo” - altos loadings en vader_comp_mean (0.40), finbert_score_mean (0.38), roberta_score_mean (0.38).
- **PC2 (16.8 % varianza):** “Volumen y negatividad” - altos loadings en vader_neg_mean (0.60), tweet_count (0.46).
- **PC3 (12.5 % varianza):** “Señales FOMO/FUD” - altos loadings en fomo_count_mean (0.70), fud_count_mean (0.58).

4.4. Resultados de modelado: Random Forest

Cuadro 6: Resultados Random Forest - MAE horizonte 24h (Walk-Forward CV)

Escenario	MAE	vs BASE	Ranking
PCA_ALL	0.0586	-3.9 %	1
PCA_VADER	0.0598	-2.0 %	2
FinBERT	0.0604	-1.0 %	3
BASE	0.0610	–	4
PCA_FinBERT	0.0611	+0.2 %	5
VADER	0.0620	+1.6 %	6

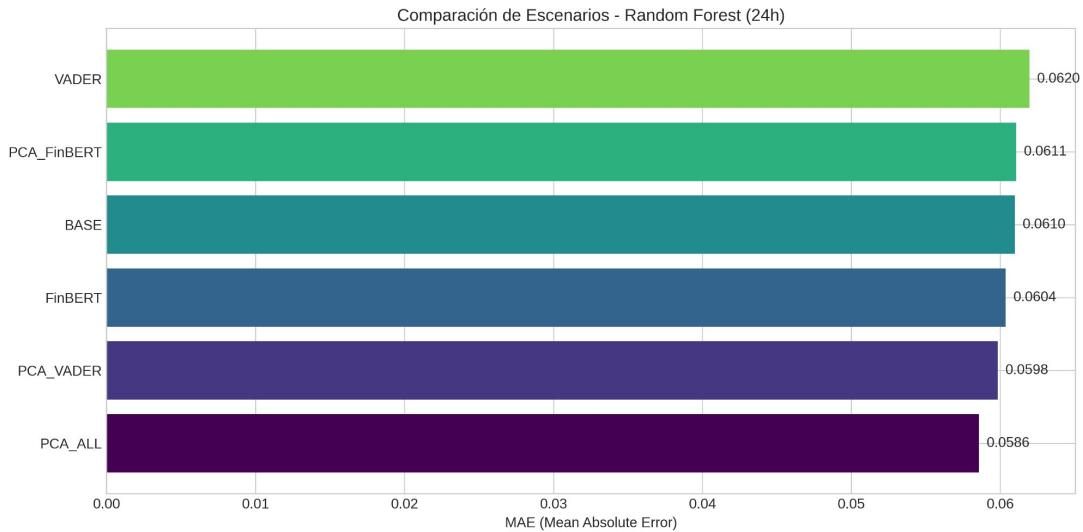


Figura 5: Comparación de MAE por escenario para Random Forest (horizonte 24h).

Para Random Forest, el escenario **PCA_ALL** obtiene el mejor desempeño, mejorando un 3.9 % respecto al baseline. Notablemente, VADER sin procesamiento empeora el desempeño (+1.6 %), mientras que FinBERT crudo y sus variantes PCA muestran mejoras marginales.

4.5. Resultados de modelado: LSTM

Cuadro 7: Resultados LSTM - horizonte 24h

Escenario	MAE	R ²	vs BASE	Ranking
PCA_FinBERT	0.0316	-0.449	-18.6 %	1
PCA_ALL	0.0381	-1.093	-1.8 %	2
BASE	0.0388	-1.242	—	3
FinBERT	0.0409	-1.634	+5.4 %	4

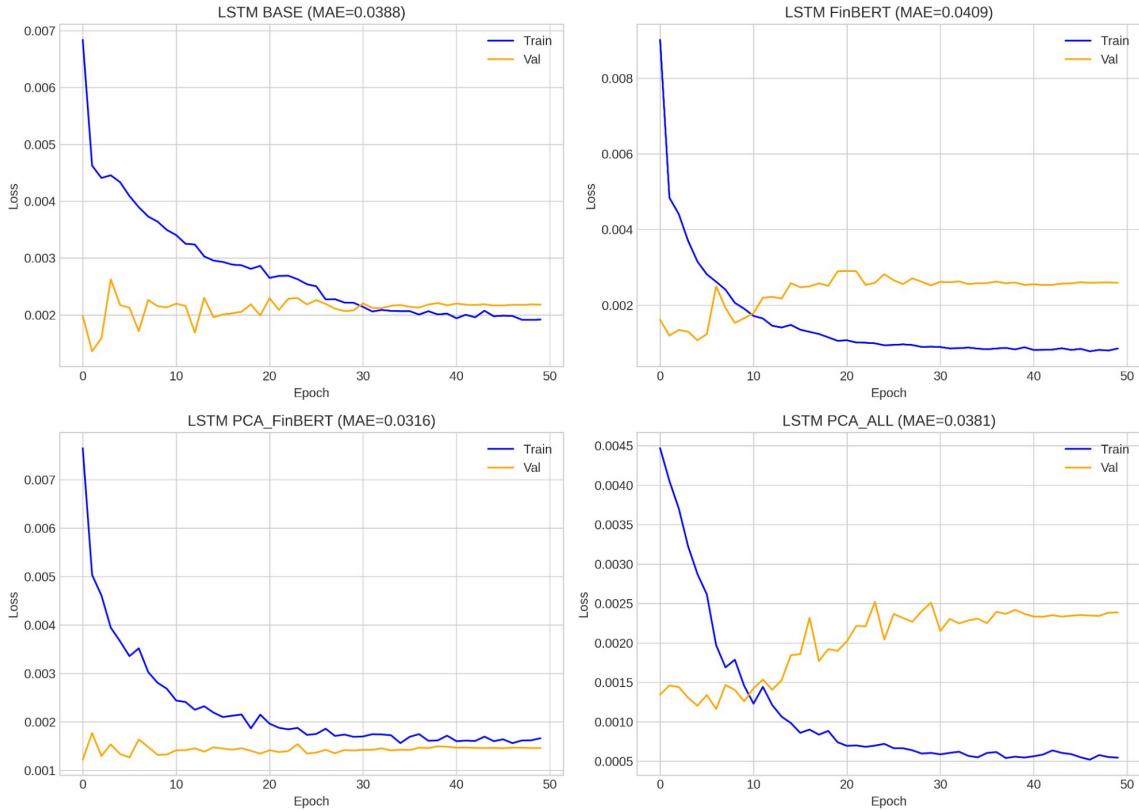


Figura 6: Curvas de aprendizaje LSTM para los cuatro escenarios evaluados. PCA_FinBERT muestra la convergencia más estable con menor validation loss.

El resultado más destacado es que **LSTM + PCA_FinBERT** logra una mejora del **18.6 %** en MAE respecto al modelo base. Este escenario también presenta el mejor R² (-0.449 vs -1.242 del baseline) y las curvas de aprendizaje más estables (Figura 6).

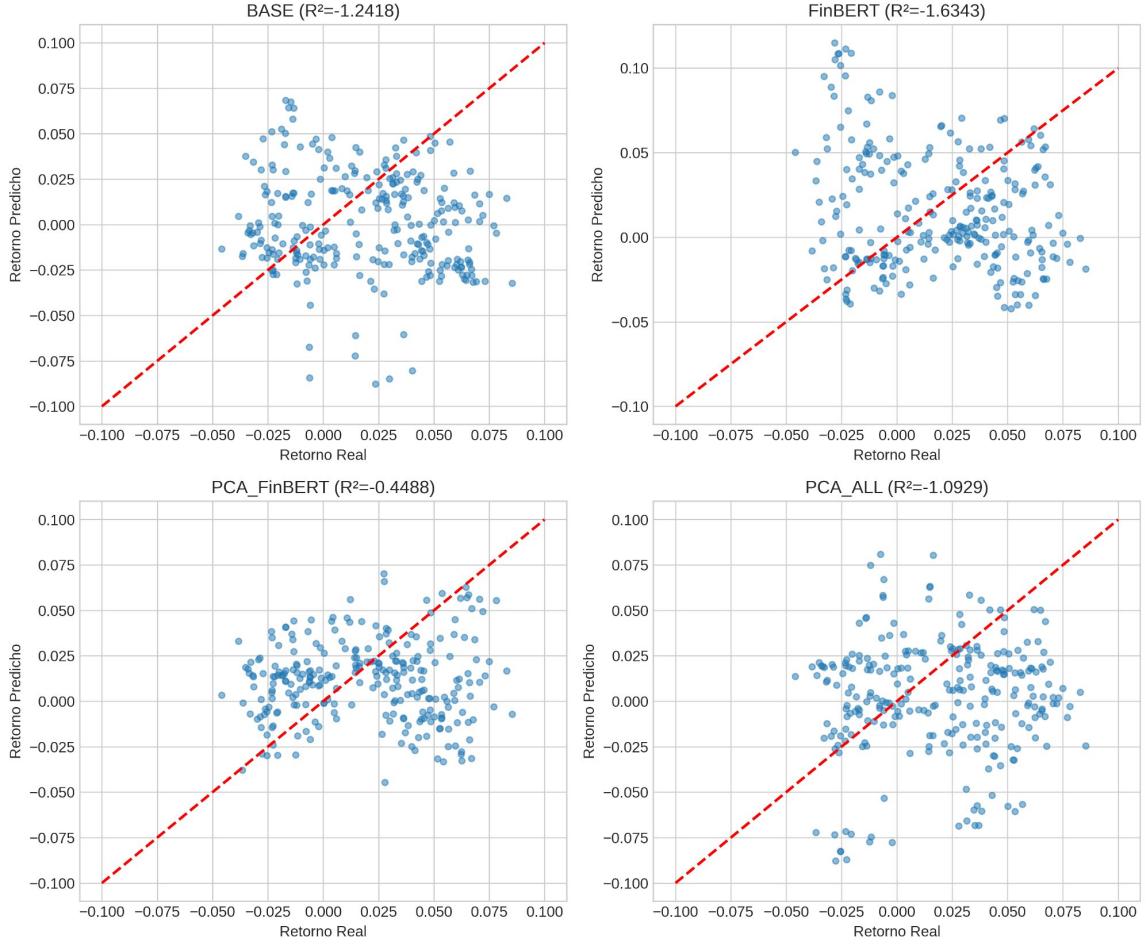


Figura 7: Scatter plots de predicciones vs retornos reales para cada escenario LSTM.

Paradójicamente, FinBERT crudo (sin PCA) empeora el desempeño del LSTM (+5.4 %), sugiriendo que el modelo sufre de overfitting cuando se incluyen demasiadas features de sentimiento correlacionadas. El PCA actúa como un regularizador efectivo al reducir la dimensionalidad y eliminar redundancia.

4.6. Resumen comparativo

Cuadro 8: Resumen de mejores modelos por familia

Familia	Mejor Escenario	MAE	Mejora vs BASE
Naive	—	0.0380	—
Ridge	BASE	0.0556	—
Random Forest	PCA_ALL	0.0586	-3.9 %
LSTM	PCA_FinBERT	0.0316	-18.6 %

5. Discusión

5.1. Efectividad de FinBERT vs VADER

Los resultados confirman la superioridad de FinBERT sobre VADER para capturar la relación entre sentimiento social y retornos de Bitcoin. La correlación 4.2 veces más fuerte ($\rho = 0.113$

vs $\rho = 0,027$) se explica por la capacidad de FinBERT para comprender el contexto financiero específico de los tweets sobre criptomonedas.

Sin embargo, una correlación más fuerte no garantiza mejor predicción. FinBERT crudo empeora el LSTM debido a la inclusión de múltiples features correlacionadas que introducen ruido y favorecen el overfitting.

5.2. Rol del PCA como regularizador

El hallazgo más importante de este estudio es el efecto regularizador del PCA. Al reducir las 8-15 features de sentimiento a 3 componentes principales:

1. Se elimina la multicolinealidad entre variables de sentimiento.
2. Se captura la información más relevante (77-84 % de varianza).
3. Se reduce el riesgo de overfitting en modelos complejos como LSTM.

Esto explica por qué PCA_FinBERT supera dramáticamente a FinBERT crudo en LSTM (-18.6 % vs +5.4 %).

5.3. Interpretación de los componentes principales

El análisis de loadings revela que los componentes principales tienen interpretación semántica clara:

- **PC1:** Representa el “consenso de sentimiento positivo” entre todos los métodos, útil para capturar optimismo/pesimismo generalizado.
- **PC2:** Captura la actividad (volumen de tweets) y sentimiento negativo, potencialmente relacionado con períodos de pánico.
- **PC3:** Aísla las señales FOMO/FUD específicas del dominio crypto, con correlación negativa significativa con retornos ($\rho = -0,098$).

La correlación negativa de PC3 sugiere que períodos de alto FOMO/FUD preceden movimientos adversos, consistente con el comportamiento contrariante del mercado.

5.4. Implicaciones para la hipótesis de eficiencia de mercado

A pesar de las mejoras obtenidas, todos los modelos presentan R^2 negativo, indicando que predecir la media sería más preciso en términos de varianza explicada. Esto es consistente con la hipótesis de eficiencia de mercado en su forma semi-fuerte: aunque existe información predictiva en el sentimiento social, su explotación sistemática es limitada.

5.5. Limitaciones

- **Período limitado:** 6 meses que incluyen el crash de mayo 2021, potencialmente atípico.
- **Fuente única:** Solo Twitter como fuente social (no Reddit, Telegram, etc.).
- **Latencia:** El análisis no considera el tiempo de procesamiento en escenarios de trading real.
- **Costos de transacción:** No se evalúa la rentabilidad neta de estrategias de trading.

6. Conclusiones

Este estudio evaluó sistemáticamente la capacidad predictiva del sentimiento social sobre el precio de Bitcoin, comparando métodos tradicionales con modelos de NLP especializados y técnicas de reducción de dimensionalidad. Las principales conclusiones son:

1. **FinBERT supera a VADER en correlación:** $\rho = 0,113$ ($p < 0,001$) vs $\rho = 0,027$ (no significativo), confirmando que modelos especializados en finanzas capturan mejor la relación sentimiento-precio.
2. **PCA mejora significativamente la predicción:** La combinación LSTM + PCA_FinBERT logra MAE = 0.0316, mejorando un 18.6 % respecto al baseline técnico.
3. **PCA actúa como regularizador:** Reduce overfitting al eliminar redundancia entre features de sentimiento correlacionadas.
4. **El PCA combinado es óptimo para Random Forest:** PCA_ALL (combinando VADER, FinBERT y RoBERTa) obtiene el mejor MAE (0.0586) para modelos de ensamble.
5. **Los componentes principales tienen interpretación semántica:** PC1 captura sentimiento general, PC2 volumen/negatividad, PC3 señales FOMO/FUD.

6.1. Implicaciones prácticas

Para investigadores y practitioners interesados en incorporar sentimiento social en modelos de predicción de criptomonedas, recomendamos:

- Utilizar FinBERT sobre VADER para extracción de sentimiento financiero.
- Aplicar PCA para reducir dimensionalidad antes de entrenar modelos complejos.
- Considerar LSTM con features PCA para capturar dependencias temporales.
- Interpretar los componentes principales para obtener insights sobre el mercado.

6.2. Trabajo futuro

- Incorporar múltiples fuentes sociales (Reddit, Telegram, Discord).
- Evaluar arquitecturas Transformer end-to-end (sin extracción de features).
- Analizar períodos más extensos incluyendo diferentes regímenes de mercado.
- Implementar backtesting con costos de transacción realistas.
- Explorar técnicas de ensemble que combinen PCA con selección de features.

Agradecimientos

Los autores agradecen a la Universidad de Las Américas por el apoyo en la realización de este trabajo, y a la comunidad de Kaggle por proporcionar los datos de tweets utilizados en el estudio.

Referencias

- [1] Kristoufek, L. (2013). Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3, 3415.
- [2] Garcia, D., & Schweitzer, F. (2015). Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science*, 2(9), 150288.
- [3] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.
- [4] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.
- [5] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of ICWSM 2014*.
- [6] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*.
- [8] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- [9] Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Proceedings of EMNLP 2020*.
- [10] Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [11] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.
- [12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- [13] Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. *CS229 Stanford University*.
- [14] Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, 1(3), 1.
- [15] Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, 65, 101188.
- [16] Mohapatra, S., et al. (2022). Cryptocurrency price analysis using sentiment analysis of Twitter data. *Applied Sciences*, 12(4), 2180.
- [17] Huang, X., et al. (2019). FinBERT: A Pre-trained Language Model for Financial Communications. *SSRN Electronic Journal*.
- [18] Bouteska, A., & Mefteh-Wali, S. (2022). The predictive power of investor sentiment for Bitcoin returns. *The Social Science Journal*.