Memoria EDA: SP500 y Factores que Influyen en su Evolución

Definición del Sentimiento

El sentimiento en redes sociales refleja la actitud o percepción que tienen los usuarios sobre el mercado en un momento dado. En este proyecto, se usa como una medida aproximada del "estado de ánimo" de los inversores, expresado a través de posts, titulares y comentarios. Analizar este sentimiento permite detectar momentos de euforia o miedo que podrían influir, o incluso anticipar, movimientos en el SP500.

Obtención de Datos

Reddit:

- **API oficial**: Se recopilaron cerca de **40.000 publicaciones**.
- **Web scraping con API**: Se obtuvieron **433.695 registros**.
- **Kaggle**: Se utilizaron datasets con **1.900.905 publicaciones**.

Todos estos datos provienen de subreddits relacionados con el mercado bursátil, tales como *WallStreetBets*, *SP500* y *Stock Market*.

Noticias:

Debido a la dificultad de obtener noticias mediante scraping (por restricciones legales, técnicas o APIs de pago), se optó por usar datasets extraídos de Kaggle. Se recopilaron titulares y fechas de medios como **Reuters, The Guardian, CNBC** y un dataset completo de **ABC News** desde 2003 hasta 2021.

Datos Históricos del SP500:

Importados desde la **API gratuita de Yahoo Finance**. Este dataset se presentó como un *MultiIndex*, lo que generó problemas al hacer merges con otros datasets de diferente formato (arrays planos), especialmente al graficar el precio junto con el sentimiento.

Indicadores Económicos:

- CSV descargado de Kaggle: **ForexFactoryData**
- Datos obtenidos mediante la **API gratuita de FRED**

Datos COT (Commitment of Traders):

Descargados directamente desde la página oficial de la **CFTC (Commodity Futures Trading Commission)**, desde 2011 hasta 2025. Se unificaron y almacenaron en un único archivo CSV.

Procesamiento de Datos

Los datasets fueron limpiados, renombrados y unidos varias veces hasta eliminar completamente valores nulos o vacíos. Un caso particular fue el de Reddit, donde se observó un **pico de publicaciones en 2021**, coincidiendo con el fenómeno de **Gamestop y los meme stocks**, lo cual se refleja en la actividad de *WallStreetBets*.

En el caso de los datasets de noticias, se estandarizaron columnas como `date` y `publish_date` para permitir su integración.

Los datos de COT, inicialmente en formato `.txt`, se procesaron y consolidaron en un solo dataframe CSV.

Además, se aplicó una **limpieza textual avanzada** a los posts de Reddit para eliminar ruido como enlaces, menciones o imágenes, que interferían en la correcta interpretación del análisis de sentimiento.

Análisis de Sentimiento

Se evaluaron tres herramientas: **VADER**, **TextBlob** y **BERT**. Finalmente se eligió **VADER** por su alta precisión en textos de redes sociales y su bajo tiempo de procesamiento. A modo de comparación:

- **BERT**: ~2 horas para analizar 433.000 posts.
- **VADER**: <10 minutos para la misma cantidad.

Aunque BERT ofrecía mayor profundidad, **VADER resultó ser la mejor opción** considerando recursos y tiempo, sin comprometer significativamente la calidad del análisis.

Cambio de Hipótesis

Inicialmente, el objetivo era analizar **el impacto del sentimiento de redes sociales en el precio del SP500**. Sin embargo, dada la amplitud de datos recopilados y la situación reciente del mercado, la hipótesis evolucionó a:

> **"¿Qué factores influyen en el precio del SP500?"**

Esto permitió realizar un análisis más completo y valioso.

Objetivos del Análisis

- **Redes Sociales y Noticias**: Evaluar si el sentimiento puede predecir o reaccionar al movimiento del SP500.
- **Indicadores Económicos**: Examinar la correlación entre la economía real (PIB, inflación, empleo, tasas de interés, liquidez) y el índice bursátil.
- **Datos COT**: Analizar si las posiciones de los *Asset Managers* (instituciones especuladores) anticipan cambios relevantes en el SP500.
- **Tendencia Estacional**: Observar patrones históricos mensuales en el SP500, destacando periodos con mejor o peor comportamiento.

Presentación

Se diseñaron gráficos de líneas simples para facilitar la comprensión visual del análisis. Se evitó usar visualizaciones complejas, priorizando la claridad. También se incluyeron algunos **memes encontrados en Reddit**, con el objetivo de captar la atención del público joven.

Problemas Encontrados

- **Twitter/X**: Se intentó analizar esta red, pero los límites de su API y sus estrictas políticas de scraping lo hicieron inviable. Librerías como *Tweepy* y *BeautifulSoup* no fueron efectivas.
- Se buscó un dataset de tweets durante la pandemia para evaluar el sentimiento en ese periodo, pero el proceso requería “hidratar” los datos vía la API oficial, lo cual también

presentó limitaciones.

Por estos motivos, **se optó por Reddit**, que es más flexible en términos de scraping y API.

Conclusión

El análisis de sentimiento mostró correlaciones significativas, especialmente durante periodos de incertidumbre (pandemia, crisis, etc.). Además, los datos económicos y los reportes COT respaldaron la hipótesis revisada. Las variables analizadas **no actúan de forma aislada**, sino que **se retroalimentan entre sí**, formando un ecosistema de influencia sobre el SP500.