Juan Diego Sánchez / Fundamentos de Data Science

2025-04-14

Proyecto Final

Análisis del Impacto del Sentimiento en Twitter sobre el Movimiento Diario de la Acción de Tesla (TSLA)

Objetivos

- Evaluar si los tweets de Elon Musk tienen impacto en la variación diaria del precio de la acción de Tesla (TSLA)
- Construir un modelo predictivo que determine si la acción sube o baja utiizando análisis de sentimiento y métricas de difusión

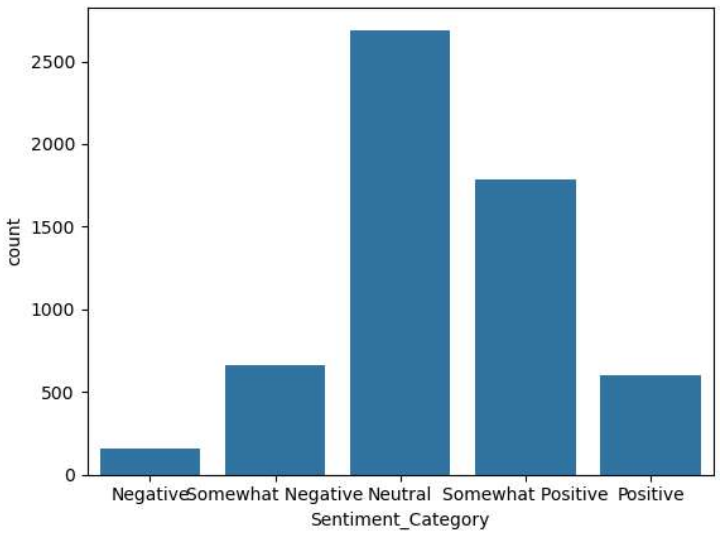
Entendimiento de los Datos

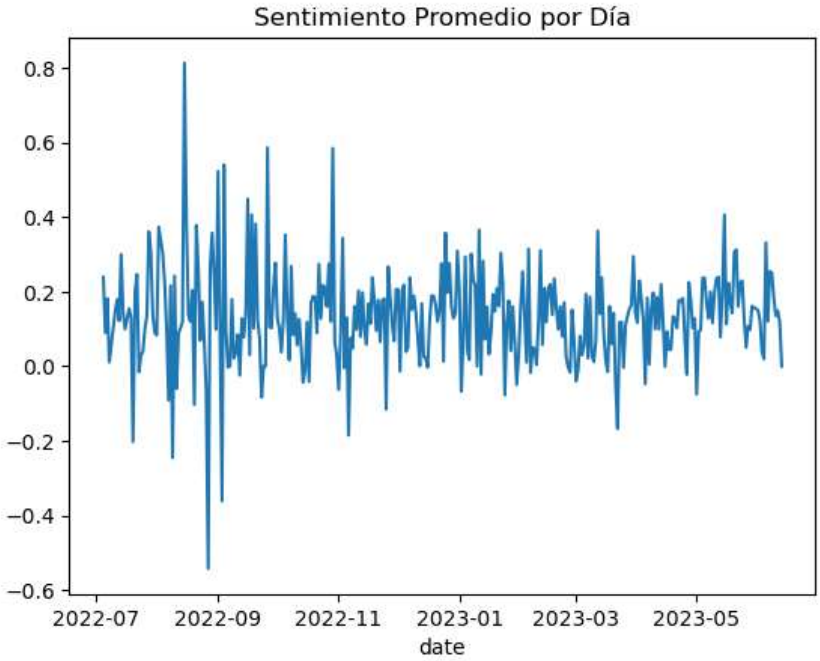
Se trabajará con un dataset publicado en kaggle.com de la cual se extraerá el texto del tweet, fecha de publicación para luego ser procesados y clasificados en función de un análisis de sentimiento. Por el dato del mercado de valores se extraerá el valor de la acción en USD al cierre junto con el volumen de transacciones realizadas durante el mismo periodo de tiempo

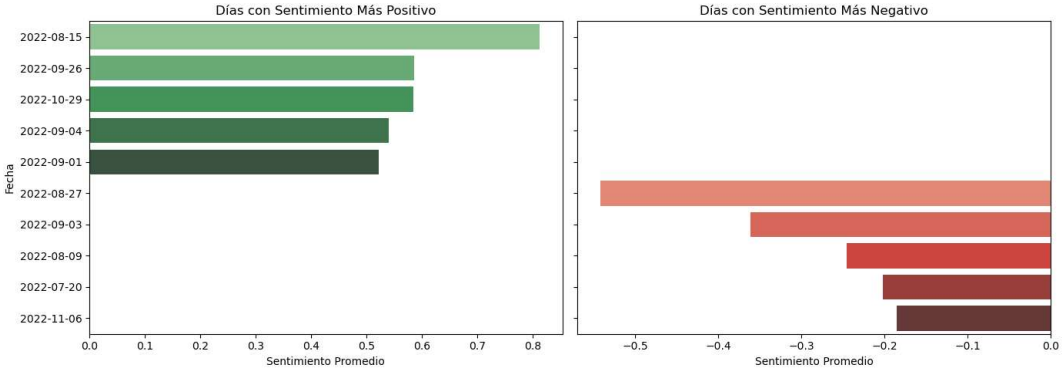
Se inicia la preparación de los datos revisando si existen valores duplicados o vacíos. Se concluye que dentro de las características a ser evaluadas no existen datos que puedan afectar el análisis. A posterior, se ajusta el formato de la columna fecha a datatime.

En el caso del dataset de tweets se procede a homogeneizar el contenido del texto; se elimina los URLs, menciones, hashtags, signos de puntuación, se convierte todo a minúsculas y se eliminan todos los pronombres que a posterior podrían llegar a desdibujar el análisis. De igual modo, se crean nuevas filas con el año, mes y día para posteriores análisis. Se calcula la intensidad de sentimiento obteniendo un número que oscila del -1 a 1 que será estandarizado en clasificadores, e.g: negativo, parciamente negativo, neutral, parcialmente positivo y positivo.

Se realizan graficas para poder observar tendencias y entender como está distribuido el sentimiento

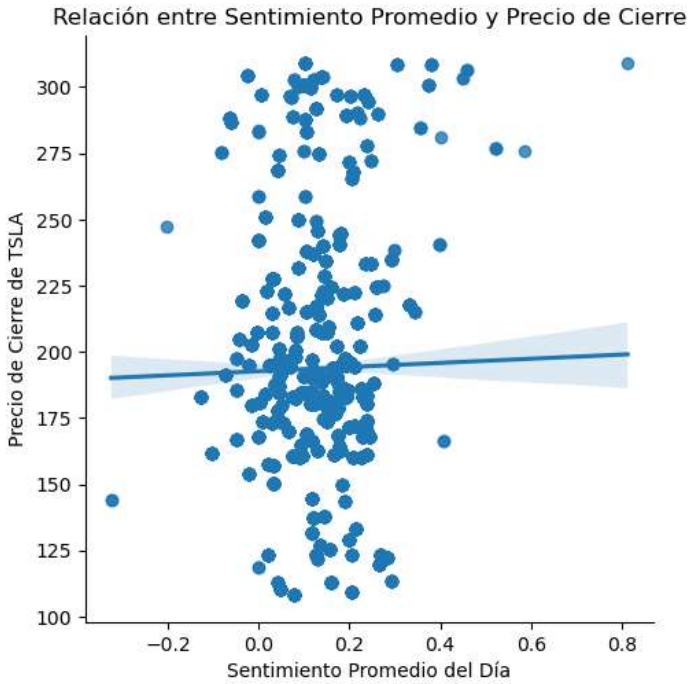






Finalmente, se realiza un merge para juntar ambos datasets usando como llave la fecha. A partir de lo cual se procede a realizar graficas para poder analizar de mejor manera el dataset, se empieza por generar un diagrama de líneas en el mismo gráfico, pero distintos ejes el sentimiento y el precio de cierre, también se gráfica el volumen diario para buscar correlaciones entre estas variables. Se gráfica de igual manera la distribución de sentimientos agrupada por mes, un scatterplot relacionando el sentimiento y el cambio porcentual en el precio y un boxplot distribuyendo el sentimiento según el volumen previamente categorizado usando la mediana como umbral.

En primer lugar, se busca una correlación entre el sentimiento promedio y el precio de cierre, para lo cual se agrupa el sentimiento por día y se compara contra el precio de cierre, se calcula la correlación obteniendo un 0,0159 lo cual se confirma con el siguiente grafico

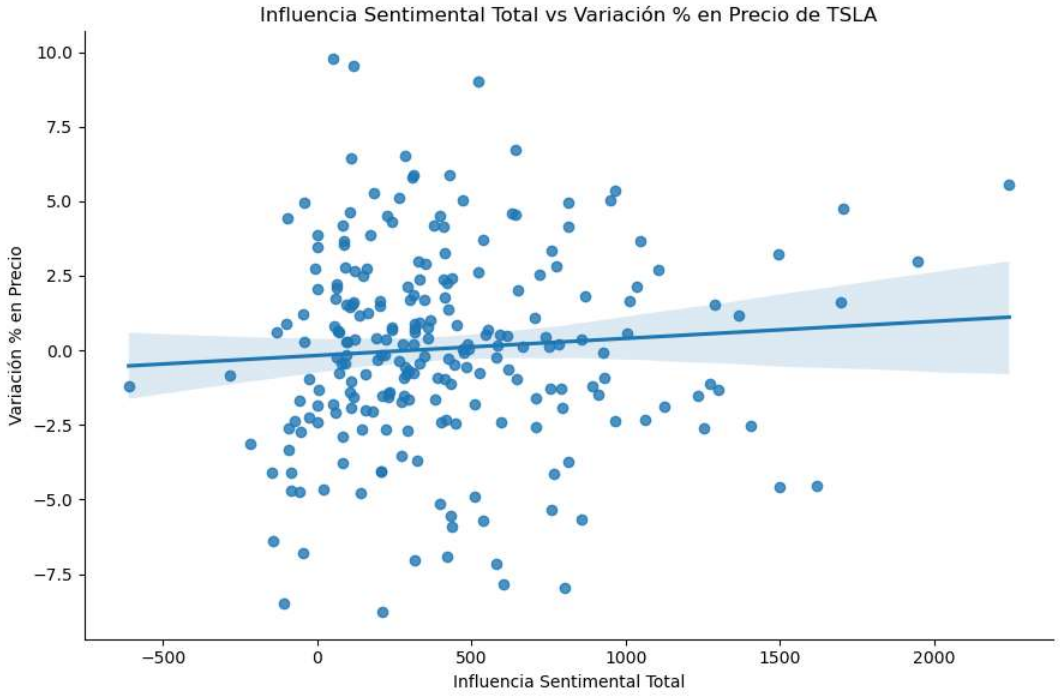


Posteriormente, se buscó validar la correlación, pero con el sentimiento promedio del día anterior, alcanzado una correlación de 0,0162.

Luego, al no alcanzar una correlación lo suficientemente significativa se procede a generar dos nuevas características que son la variación diaria del precio de la acción tanto en dólares como en porcentaje, eso se lleva a cabo calculando la resta del precio de cierre con el precio de apertura. Aquí se alcanza una correlación de 0,1532. Se prueba de igual manera la correlación contra 1 dia de desfase y con la media móvil de 3 días alcanzando 0,141 y 0,149 respectivamente.

A continuación se busca si la impacto en los días donde el sentimiento fue mas extremo produce un mayor impacto en la variación del precio para hacer esto se filtra solo los días con sentimiento promedio mayor a 0,5. En este escenario se alcanza una correlación muy alta (0,928) pero con un numero muy pequeño de muestras por lo que no podrían ser utilizadas para un modelo.

Posteriormente y bajo el criterio de ponderar el sentimiento con la amplificación pública se crea una característica adicional ponderando la cantidad de veces que fue retweeteado una publicación junto con el sentimiento cuantificado. Se procede de igual manera con las características de seguidores, favoritos y finalmente se cuantifica la influencia total juntando todas las variables anteriores.



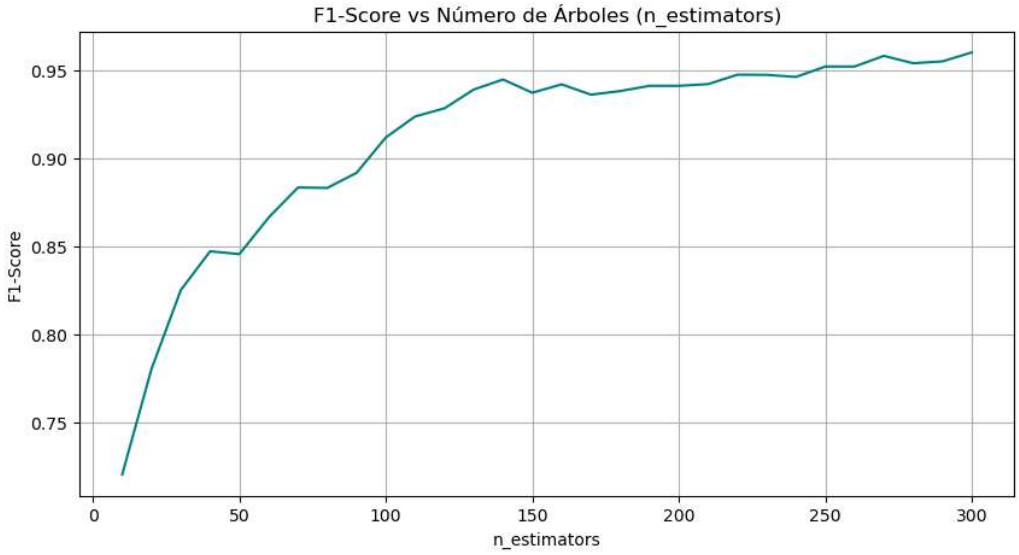
Modelos Evaluados

Se empieza el análisis con un random forest classifier alcanzando un f1 de 0,91 y un f1 de validación cruzada de 0,525. También se cuantifica la importancia de las características y se grafica concluyendo que las más importantes son el sentimiento promedio, sentimiento con lag de 1 día y sentimiento promedio de 3 días. Es decir, el modelo se esta enfocando en el sentimiento diario y la tendencia de sentimiento colectivo.

Luego, se procede a comparar la regresión lineal, voting classifer, bagging classifier y gradient boosting classifer alcanzando f1s mostrados a continuación:

Modelo	F1-score (Test)	F1-score (CV)
Regresión Logística	0,573	0,559
Bagging Classifier	0,971	0,528
Voting Classifier (Soft)	0,956	0,526
Gradient Boosting	0,971	0,535
Random Forest	0,912	0,525

Finalmente se decide por el gradient boosting classifier y se procede a ajustar los parámetros para lo cual se toma un enfoque gráfico mostrado a continuación y se decide por n_estimators igual a 150 y learning rate de 0.2:



Plan de Implementación

- Preparación del Entorno
    - Limpieza y normalización automatizada de tweets entrantes (API twitter)
    - Acceso Continuo a Datos Bursatiles (API Financiera)
- Pipeline de toma de datos y análisis
    - Tokenización y análisis de sentimiento de tweets
    - Agrupación por día y generación de características adicionales (infl_total_ sentiment_avg)
- Predicción y Evaluación Continua
    - Aplicar y monitorear el modelo diariamente considerando una recalibración periódica
- Despliegue
    - Entorno en la nube (Google Cloud, AWS, etc**)**

Conclusiones y Recomendaciones

- Se identifica un poder predictivo limitado del sentimiento, él cual logra mejorarse con métricas de amplificación como: retweets, favoritos y seguidores

- Modelos como Gradient Boosting ofrece buenos resultados, pero debe ser regularizado para evitar overfitting y memorización de patrones
- Es posible identificar correlaciones entre sentimiento y comportamiento del precio de la acción que sean estadísticamente relevantes
- Se recomienda generar análisis con granularidades menores (cada 6 horas, cada 2 horas)
- Se requiere realizar más pruebas con las características existentes buscando minimizar la diferencia entre el f1 resultante y el f1 de validación cruzada, se puede probar incluyendo el precio de la acción del día anterior