Pipeline de Ciencia de Datos
SarcOji

¿Qué problemas se identifican en el dataset?

1. **Valores Negativos Inesperados**

   Columnas como MaxEmojiNumOccurence (Cantidad de veces que se repite un emoji, debería ser 0 o positivo) y MaxEmojiPos (característica categórica que toma valores 0, 1 y 2) presentan valores de -1.

   Problema: No se usan NaN o valores categóricos claros.

2. **Outliers**

   Los rangos de medición para el análisis de sentimiento deberían estar entre -1 y 1. Sin embargo, en algunas características de algunos registros se sale de este rango.

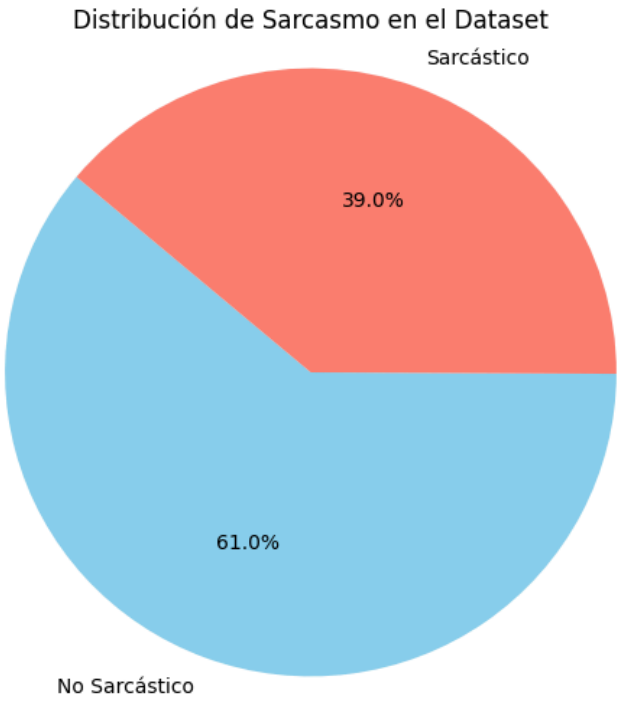
   - TextSWN: entre -3.25 y 7.75
   - EmojiVader: hasta ±3
   - EmojiSWN: hasta ±12
   - MaxEmojiNumOccurence: hasta 50

   Problema: Esto puede sesgar modelos si se comparan directamente sin normalización.

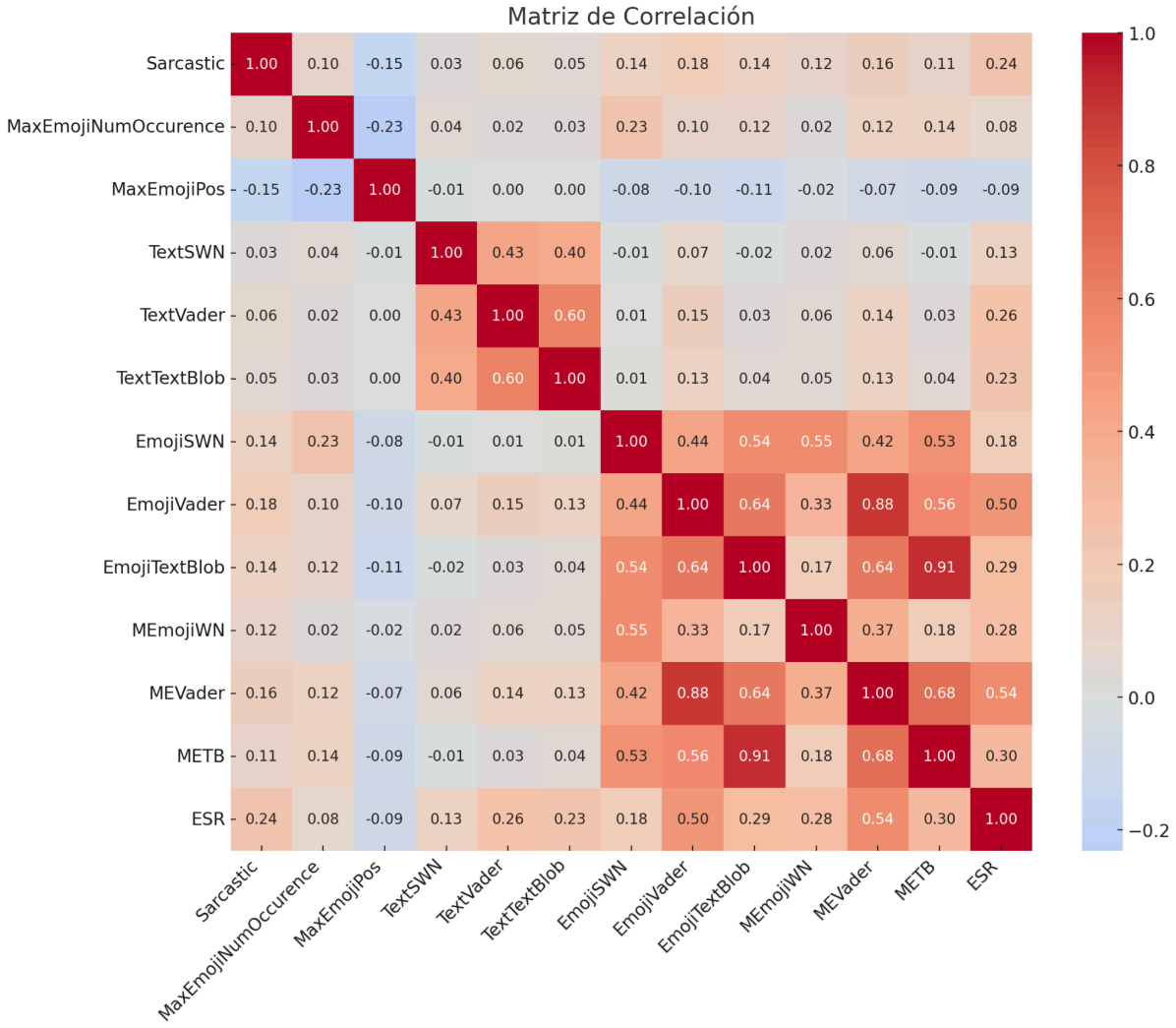
	Sarcastic	MaxEmojiNumOccurence	MaxEmojiPos	TextSWN	TextVader	TextTextBlob	EmojiSWN	EmojiVader	EmojiTextBlob	MEmojiWN	MEVader	METB	ESR
count	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000	29377.000000
mean	0.389693	1.352010	1.657045	0.100766	0.120409	0.103266	0.037296	0.103967	0.132164	-0.024083	0.071436	0.106782	0.138415
std	0.487689	1.192936	0.645701	0.495137	0.445085	0.327590	0.381404	0.413368	0.426454	0.302915	0.356319	0.401300	0.330271
min	0.000000	-1.000000	-1.000000	-3.250000	-0.994000	-1.000000	-10.000000	-2.613200	-2.750000	-10.000000	-0.830200	-1.000000	-1.000000
25%	0.000000	1.000000	2.000000	0.000000	-0.012900	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.093000
50%	0.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.221000
75%	1.000000	1.000000	2.000000	0.250000	0.476700	0.285714	0.250000	0.440400	0.225000	0.000000	0.440400	0.000000	0.456000
max	1.000000	50.000000	2.000000	7.750000	0.999300	1.000000	12.500000	3.086700	2.000000	7.500000	0.834600	1.000000	1.000000

3. **Desequilibrio en la Clase Objetivo**

- La clase está moderadamente desbalanceada (~61% no sarcástico vs ~39% sarcástico), lo que puede sesgar el resultado del modelo.



¿Qué descubrieron al analizar los datos?



1. **Redundancia entre variables de sentimiento de emojis**

- EmojiVader ↔ MEVader ≈ 0.88
- EmojiTextBlob ↔ METB ≈ 0.91

Varias columnas están fuertemente correlacionadas porque miden lo mismo desde una perspectiva ligeramente distinta.

2. **El sarcasmo no se relaciona directamente con el sentimiento**

   La variable Sarcastic muestra muy poca o ninguna correlación con los puntajes de sentimiento, tanto del texto como de los emojis:

   - Sarcastic ↔ TextVader ≈ 0.014
   - Sarcastic ↔ EmojiVader ≈ 0.035
   - Sarcastic ↔ TextBlob/ESR ≈ cercano a 0

   El sarcasmo no se expresa simplemente como “sentimiento negativo” o “positivo”.

3. **Emojis y texto no están alineados emocionalmente**

- TextVader ↔ EmojiVader ≈ 0.03
- TextTextBlob ↔ EmojiTextBlob ≈ 0.005

  El sentimiento expresado por el texto y los emojis puede divergir. En contextos sarcásticos, por ejemplo, un emoji feliz puede acompañar un comentario amargo o negativo.

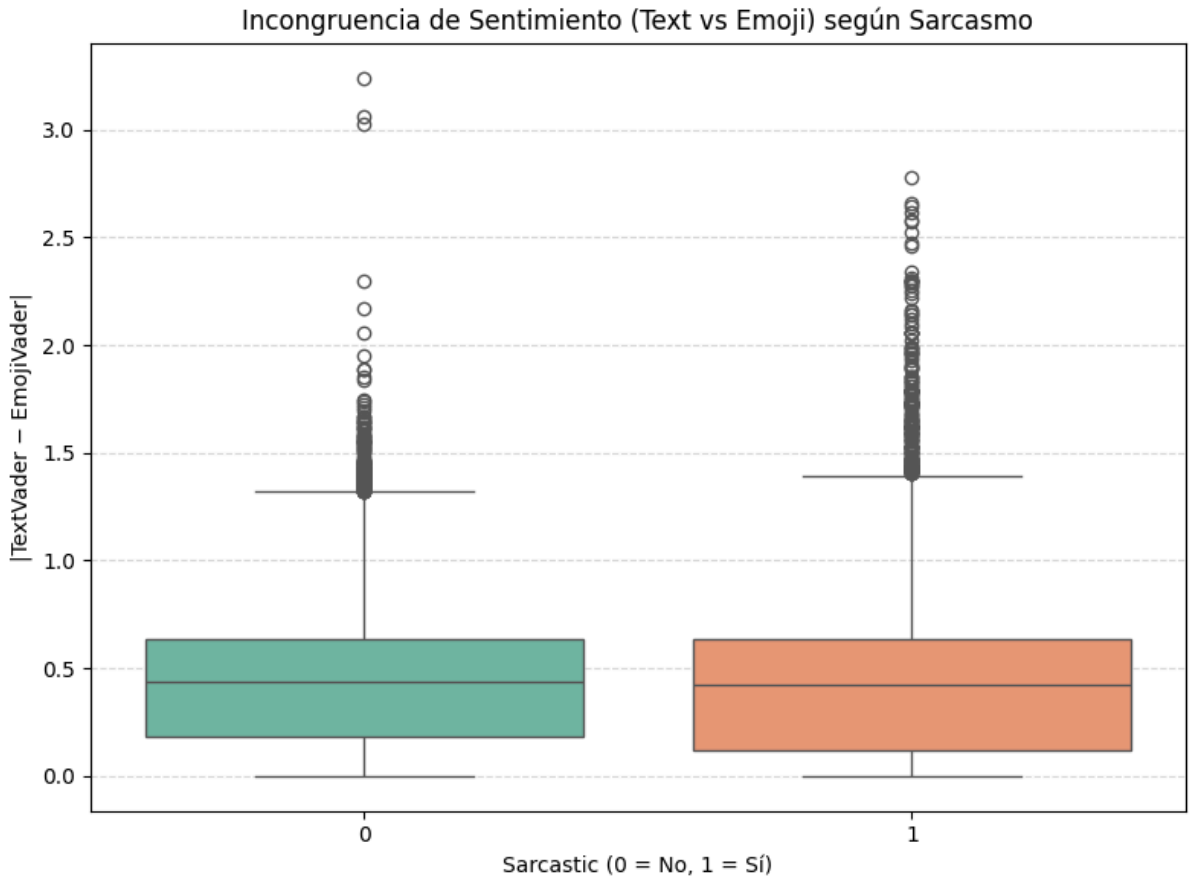
¿Qué reflejan los patrones de tendencia?

1. **Tendencia a la Incongruencia en el Sarcasmo**

   En observaciones marcadas como sarcásticas, se observa con mayor frecuencia una disonancia entre el sentimiento del texto y el de los emojis. Esto sugiere que la incongruencia emocional es un patrón característico del sarcasmo:

   - Ejemplo: texto aparentemente positivo con emojis negativos o neutros (o viceversa).

   Esta contradicción es una pista importante para identificar la ironía.



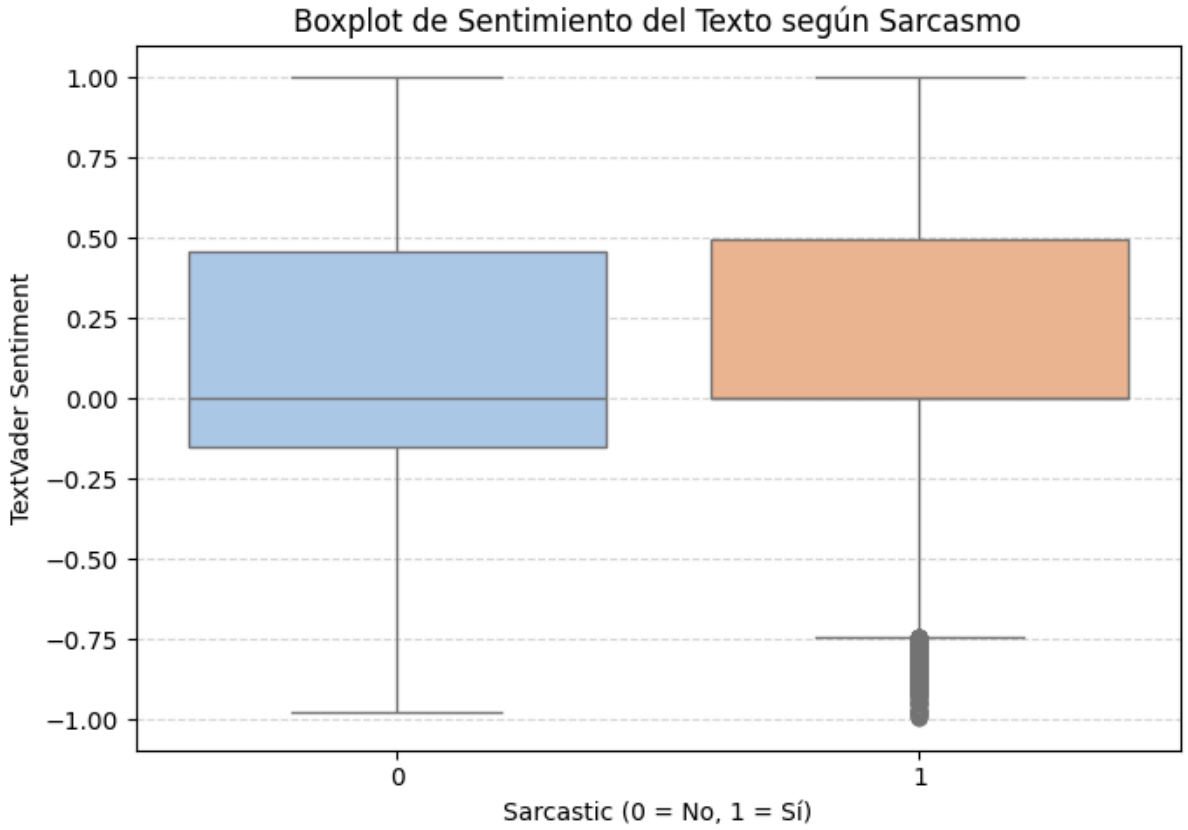
Si los casos sarcásticos tienen una incongruencia más alta, su caja (box) estará desplazada hacia arriba comparada con los no sarcásticos.

2. **Ausencia de Tendencias Lineales Simples**

Las correlaciones entre los puntajes de sentimiento (texto/emojis) y la etiqueta Sarcastic son prácticamente nulas. Esto indica que no hay una tendencia lineal evidente como:

- “Cuanto más negativo sea el texto, más probable es que sea sarcástico.”

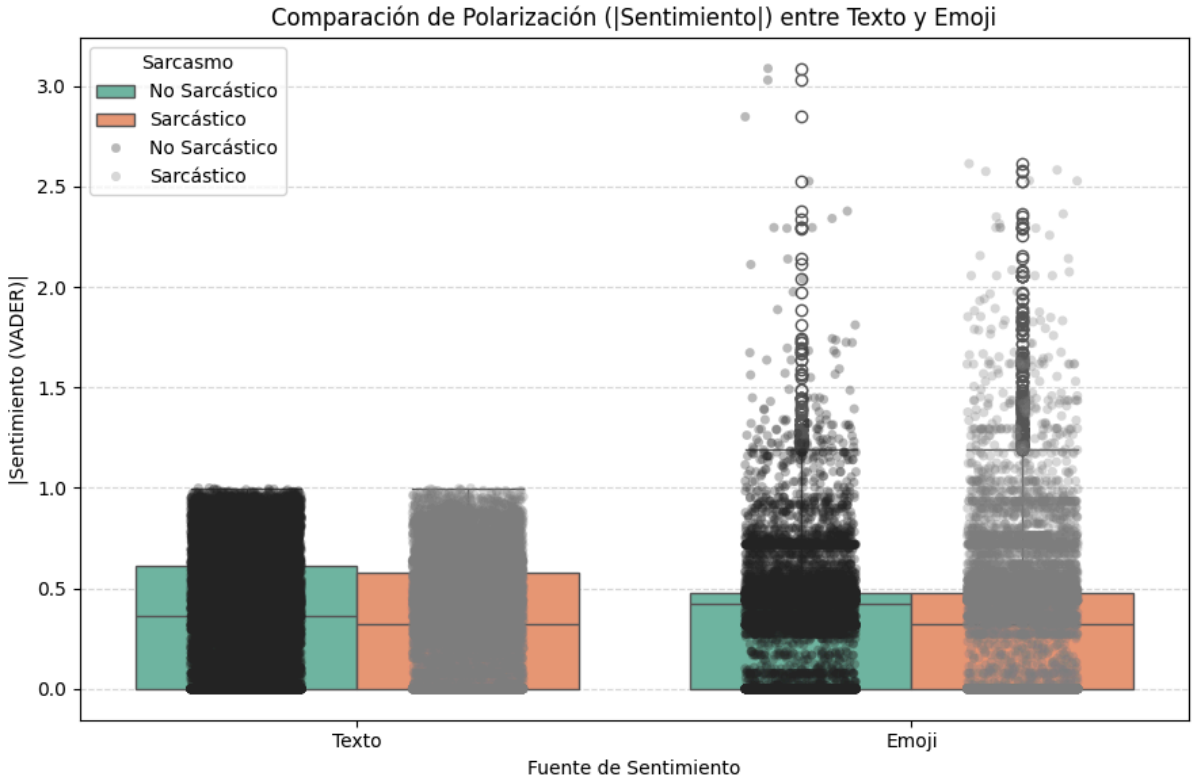
El sarcasmo no responde a una simple escala positiva-negativa. Requiere interpretar contradicciones, contexto o incluso cultura.



3. Tendencia de emojis a aportar más polarización

Las métricas de sentimiento relacionadas con emojis (`EmojiSWN`, `EmojiTextBlob`) tienen mayor rango que las de texto.

Esto sugiere que los emojis tienden a amplificar la carga emocional del contenido, lo cual puede ser usado como una pista en el sarcasmo, donde muchas veces se exagera una emoción para provocar el efecto irónico.



- Si los boxplots de los emojis están más altos (especialmente en Sarcástico), significa que los emojis tienden a expresar sentimientos más extremos que el texto.

- Si esta diferencia es especialmente marcada en los ejemplos sarcásticos refuerza la hipótesis de que los emojis amplifican la emoción en mensajes irónicos.

4. **Casos Sarcásticos Extremadamente Neutros**

   Algunos ejemplos sarcásticos tienen:

   - `TextVader ≈ 0`
   - `EmojiVader ≈ 0`

   Esto puede reflejar sarcasmo sutil, en el que ni el texto ni los emojis muestran polaridad clara. Este patrón sugiere que algunos tipos de sarcasmo no se manifiestan con emociones exageradas, sino con tono plano o seco (deadpan sarcasm). Si el sarcasmo está presente, probablemente lo está por contexto, ironía, contradicción o emojis.

Text	TextVader	EmojiVader
1 hour and 15 minute experiment after class 😊	0.0000	0.7184
Late nights, early mornings 😩 😩 😩	0.0000	-0.3987