Práctica de Laboratorio:
Análisis Exploratorio de Datos - Data Wrangling

Docente: Ana Maria CuadrosValdivia

SarcOji

Contexto

El uso de sarcasmo en redes sociales es común y desafiante de detectar automáticamente, debido a su naturaleza implícita, contextual y subjetiva. Los emojis, como expresiones visuales del lenguaje digital, añaden matices emocionales que pueden reforzar, contradecir o suavizar el contenido textual.

El objetivo principal de SarcOji es explorar cómo los emojis actúan cómo señales de refuerzo emocional o connotativo en textos breves (como tweets o publicaciones de Facebook) y cómo estos pueden ser aprovechados por modelos para mejorar la detección del sarcasmo, una tarea desafiante por su ambigüedad contextual y semántica.

Origen

SarcOji fue compilado a partir de cinco datasets públicos de sarcasmo, provenientes de publicaciones en Twitter y Facebook.

Contenido del Dataset

- **Plataformas**: Twitter (≈82%), Facebook (≈18%)
- **Idioma**: Solo textos en inglés
- **Condición obligatoria**: Todos los textos contienen **al menos un emoji**

Estructura

Cada registro incluye:

- El texto original
- Etiqueta de sarcasmo (1 o 0)

- Emojis utilizados
- Emoji más frecuente
- Posición del emoji más frecuente
- Puntajes de sentimiento del texto y emojis, calculados con:
    - SentiWordNet
    - VADER
    - TextBlob
    - Emoji Sentiment Ranking (ESR)

1. Analiza el comportamiento de tus datos

Representación de un Registro

Un registro de SarcOji es una instancia de texto en inglés extraída de redes sociales (Twitter o Facebook), etiquetada como sarcástica (Sarcastic=1) o no sarcástica (Sarcastic=0), que contiene al menos un emoji y diversas características derivadas relacionadas con el análisis de sentimiento y uso de emojis.

- **Ejemplo de un registro de SarcOji etiquetado cómo SARCÁSTICO**


```
{
  "Text": "Oh great, another Monday 😂😂 ",
  "Sarcastic": 1,
  "Emojis": ["😂", "😂"],
  "MaxEmoji": "😂",
  "MaxEmojiNumOccurence": 2,
  "MaxEmojiPos": 2,
  "TextSWN": -0.125,
  "TextVader": 0.0,
  "TextTextBlob": -0.100000,
  "EmojiSWN": 1.125,
  "EmojiVader": -0.3987,
  "EmojiTextBlob": 0.0,
  "MEmojiWN": 1.125,
  “MEVader": -0.3987,
  "METB": 0.0,
  "ESR": -0.368
}
```

Este registro representa una entidad sarcástica, donde el texto tiene una polaridad negativa y el emoji dominante (😂) tiene una polaridad positiva, lo que puede indicar incongruencia emocional y sarcasmo.

- **Ejemplo de un registro de SarcOji etiquetado cómo NO SARCÁSTICO**


```
{
  "Text": "6 months ago I lost my baby boy, the most precious thing in my life 😭😭😭",
  "Sarcastic": 0,
  "Emojis": ["😭", "😭", "😭"],
  "MaxEmoji": "😭",
  "MaxEmojiNumOccurence": 3,
  "MaxEmojiPos": 2,
  "TextSWN": -0.125,
  "TextVader": 0.5598,
  "TextTextBlob": 0.294444,
  "EmojiSWN": 0.000,
  "EmojiVader": -0.4767,
  "EmojiTextBlob": -0.2000,
  "MEmojiWN": 0.000,
  "MEVader": -0.4767,
  "METB": -0.2000,
  "ESR": -0.093
}
```


1. **¿Qué es un emoji "😂" hablando computacionalmente?**

   Un emoji es un **carácter Unicode** que ocupa una posición específica en el estándar de codificación de texto, y que puede ser renderizado como una imagen gráfica por el sistema operativo, navegador o aplicación.

Salida	Unicode
😂	U+1F60A
"A"	U+0041

Dimensiones

- Tiene 29,377 registros y 15 columnas (atributos)

Tamaño

- Es una cantidad adecuada para tareas de aprendizaje automático como clasificación de sarcasmo.

- En su formato original (dataframe serializado o .df) tiene un tamaño de 6 megabytes de peso. Puede cargarse completamente en memoria como un dataframe de pandas.

Duplicados

- Según la metodología descrita en el artículo, los datos fueron preprocesados y limpiados.

Tipo de datos

Columna	Descripción General	Tipo De Dato	Tipo
Text	Publicación de red social (Facebook o Twitter) con al menos un emoji.	string (Unicode)	-
Sarcastic	Etiqueta binaria que indica si el texto es sarcástico no sarcástico : 1 = Sarcastico, 0 = No Sarcástico	int (Binario)	Discreto
Emojis	Lista de todos los emojis presentes en el texto.	list[string (Unicode)]	-
MaxEmoji	El emoji que aparece con mayor frecuencia en el texto.	String (Unicode)	-
MaxEmojiNumOccurence	Número total de veces que el MaxEmoji aparece en el texto.	int	Discreto
MaxEmojiPos	Posición del emoji en el texto donde: 0 = Inicio, 1 = Medio, 2 = Final	int (Categórico)	Discreto
TextSWN	Puntaje de sentimiento del texto según SentiWordNet. Valores entre -1 y 1 (negativo a positivo).	float	Continuo
TextVader	Puntaje de sentimiento del texto obtenido mediante VADER, un analizador diseñado para redes sociales.	float	Continuo

TextTextBlob	Puntaje de sentimiento del texto calculado con TextBlob, una librería de procesamiento de lenguaje natural.	float	Continuo
EmojiSWN	Puntaje combinado de todos los emojis en el texto según SentiWordNet, aplicando repetición proporcional.	float	Continuo
EmojiVader	Puntaje combinado de todos los emojis usando VADER. Se usa su representación textual (demojized).	float	Continuo
EmojiTextBlob	Puntaje combinado de todos los emojis según TextBlob. Igual que los anteriores, usa descripciones de emojis.	float	Continuo
MEmojiWN	Sentimiento del MaxEmoji calculado con SentiWordNet, aplicando su descripción textual proporcional a su frecuencia.	float	Continuo
MEVader	Puntaje del MaxEmoji usando VADER sobre su texto representativo.	float	Continuo
METB	Puntaje del MaxEmoji según TextBlob, con su descripción textual expandida por intensidad.	float	Continuo
ESR	Puntaje del MaxEmoji basado en Emoji Sentiment Ranking (ESR), una base de datos que asocia emojis con polaridad emocional.	float	Continuo

Rangos de los datos (Encontradas dentro del dataset))

- ***Variables Numéricas Discretas***

Columna	Min	Max
MaxEmojiNumOccurence	-1	50
MaxEmojiPos	-1	2

- ***Variables Numéricas Continuas***

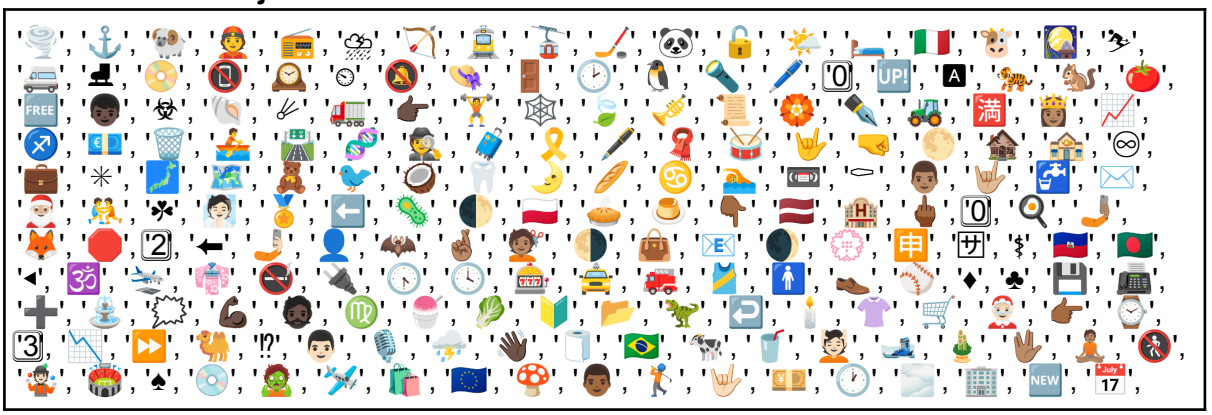
Columnas	Min	Max
TextSWN	-3.25	7.75
TextVader	-0.9940	0.9993
TextTextBlob	-1.0	1.0
EmojiSWN	-10.0	12.5
EmojiVader	-2.6132	3.0867
EmojiTextBlob	-2.75	2.0
MEmojiWN	-10.0	7.5
MEVader	-0.8302	0.8346
METB	-1.0	1.0
ESR	-1.0	1.0

- ***Variables Categóricas***

Columna	Categorías
Sarcastic	0 = No Sarcasmo, 1 = Sarcasmo
MaxEmojiPos	0 = Inicio, 1 = Medio, 2 = Final

- ***Variables Únicas***

  - **Emojis**



Problemas con el Formato

- Algunos emojis podrían estar en formatos Unicode no estándar.

Unidades de Medida

- **Puntaje de sentimiento**
  Todas las columnas de puntaje de sentimiento (TextSWN, EmojiVader, MEVader, etc.) manejan valores numéricos en el rango aproximado de: [-1, 1]

  - **Rango**

Valor	Interpretación General
`-1.0`	Sentimiento extremadamente negativo. El texto o emoji expresa emociones fuertes como ira, tristeza, frustración, burla hostil.
`0.0`	Sentimiento neutro o ambiguo. El texto o emoji no expresa una emoción clara o está emocionalmente balanceado. Puede ser difícil detectar sarcasmo desde este punto.
`1.0`	Sentimiento extremadamente positivo. El texto o emoji expresa emociones como alegría, entusiasmo, aprecio o gratitud sincera.

  - **Interpretación**

Columna	Fuente	¿Qué representa?
`TextSWN`, `EmojiSWN`, `MEmojiWN`	SentiWordNet	Sentimiento léxico basado en significados de palabras.
`TextVader`, `EmojiVader`, `MEVader`	VADER	Basado en reglas y puntuaciones léxicas orientadas a redes sociales.

		Tiene buena sensibilidad a la puntuación y modismos.
`TextTextBlob`, `EmojiTextBlob`, `METB`	TextBlob	Basado en un enfoque léxico simple. Positivo/negativo según la polaridad global del texto.
`ESR`	Emoji Sentiment Ranking	Puntaje medio de sentimiento basado en uso real de emojis en tweets anotados. Solo para emojis.

- **Ejemplo**


```
{
  "Text": "Oh great, another Monday 😂",
  "TextVader": -0.4,
  "MEVader": 0.7
}
```


Interpretación:

➢ El texto expresa un sentimiento negativo (posiblemente sarcasmo).

➢ El emoji "😂" tiene un sentimiento positivo, lo que sugiere incongruencia entre el contenido y la emoción visual, una pista común de sarcasmo.

Granularidad

- No, el dataset no contiene granularidad espacial ni temporal.

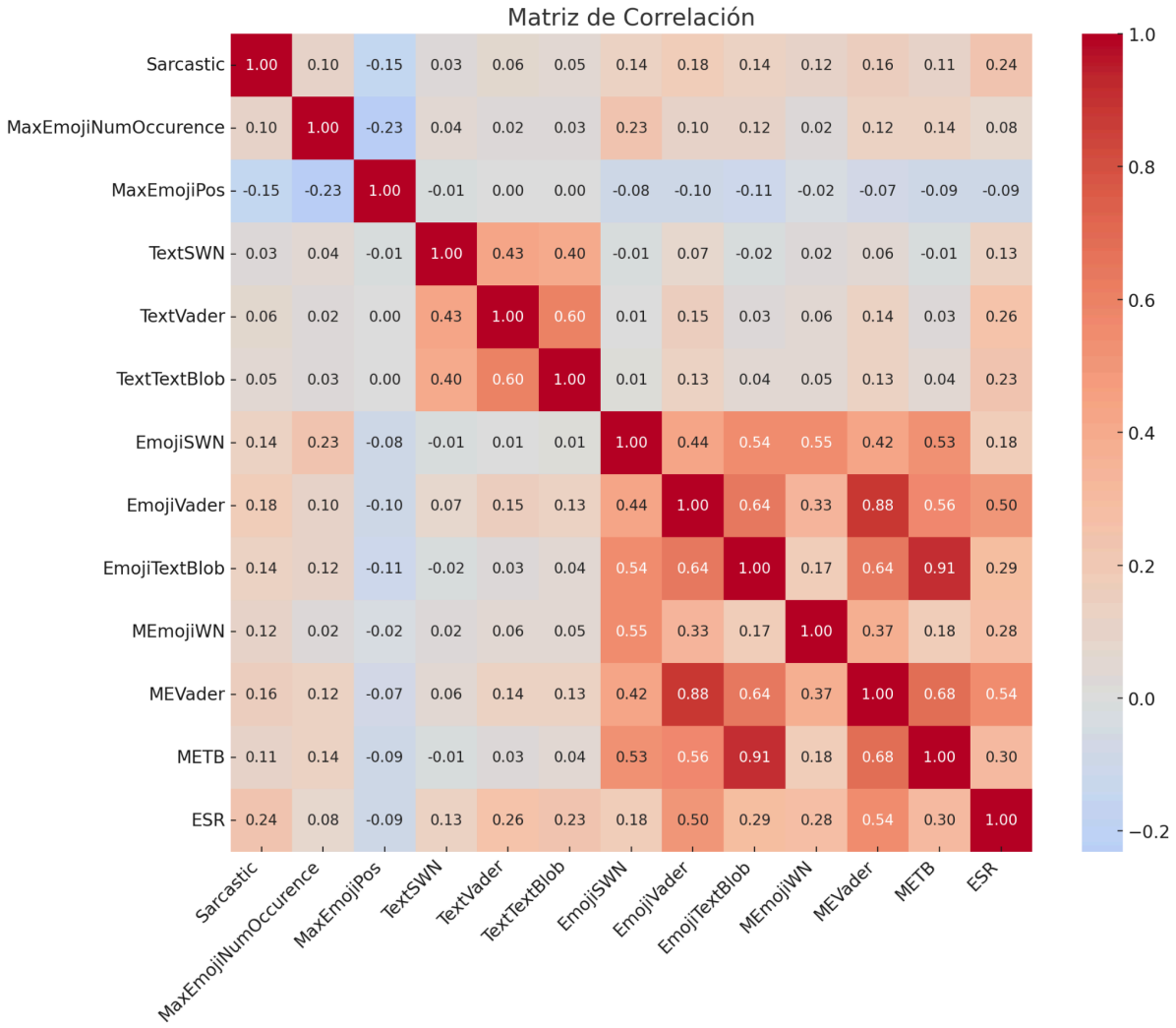
Valores Nulos

- Según la descripción del paper, el dataset fue preprocesado y curado, por lo que no presenta valores nulos.

Descripción General

	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

Correlación entre columnas



La matriz de covarianza muestra cómo varían conjuntamente las variables. Los más destacados son:

Correlación Alta:

MEVader ↔ **EmojiVader**: coeficiente ≈ 0.88

- Fuerte correlación entre el puntaje de sentimiento del emoji más representativo (MaxEmoji) y el puntaje agregado de los emojis, ambos calculados con el modelo Vader.

METB ↔ EmojiTextBlob: coeficiente ≈ 0.91

- Alta correlación entre el sentimiento del MaxEmoji (usando TextBlob) y el promedio de sentimientos de todos los emojis con TextBlob.

Correlación Baja:

TextVader ↔ EmojiVader = 0.028

TextTextBlob ↔ EmojiTextBlob ≈ 0.005

TextSWN ↔ EmojiSWN ≈ -0.002

- Estas correlaciones son extremadamente bajas o incluso negativas. Lo sorprendente aquí es que, aunque uno podría esperar que el tono emocional del texto y el de los emojis usados estén alineados, en este conjunto de datos no parece ser el caso (Sarcasmo).

Sarcastic ↔ TextVader ≈ 0.014

Sarcastic ↔ EmojiVader ≈ 0.035

Sarcastic ↔ ESR ≈ 0.039

- Todos estos valores son muy bajos. Esto es interesante porque sugiere que el sarcasmo no se detecta bien usando únicamente sentimientos superficiales. Sarcasmo puede manifestarse en frases que parecen neutrales o positivas, pero en realidad implican burla o ironía.

  Ejemplo:

  > "¡Qué idea tan brillante! Quemar la pizza otra vez..."

  A simple vista (por análisis de sentimientos), esto puede parecer positivo (“brillante”), pero en realidad es irónico y negativo

  Los algoritmos como Vader o TextBlob:

  - Analizan palabras y frases de forma literal.
  - No detectan contradicciones implícitas entre contexto, tono y significado.

  Esto explica por qué los puntajes de sentimiento no están relacionados con la presencia de sarcasmo. Haciendo que el detectar sarcasmo requiere modelar el contexto, contradicción, tono o incluso conocimiento común.

2. Análisis de outliers

Columna	Rango Promedio	Min Outlier	Max Outlier
Text	-	-	-
Sarcastic	-	-	-
Emojis	-	-	-
MaxEmoji	-	-	-
MaxEmojiNumOccurence	-	-	50

MaxEmojiPos	-	-	-
TextSWN		-3.250000	7.750000
TextVader		-	-
TextTextBlob		-	-
EmojiSWN		-10.000000	12.500000
EmojiVader		-2.613200	3.086700
EmojiTextBlob		-2.750000	2.000000
MEmojiWN		-10.000000	7.500000
MEVader		-	-
METB		-	-
ESR		-	-

¿Podemos eliminarlos? ¿Es importante conservarlos?

- En tareas de detección de sarcasmo, los outliers no necesariamente son errores, sino expresiones intensificadas.

- Para modelos robustos (transformers), puede no ser necesario, ya que esos modelos manejan bien variaciones.

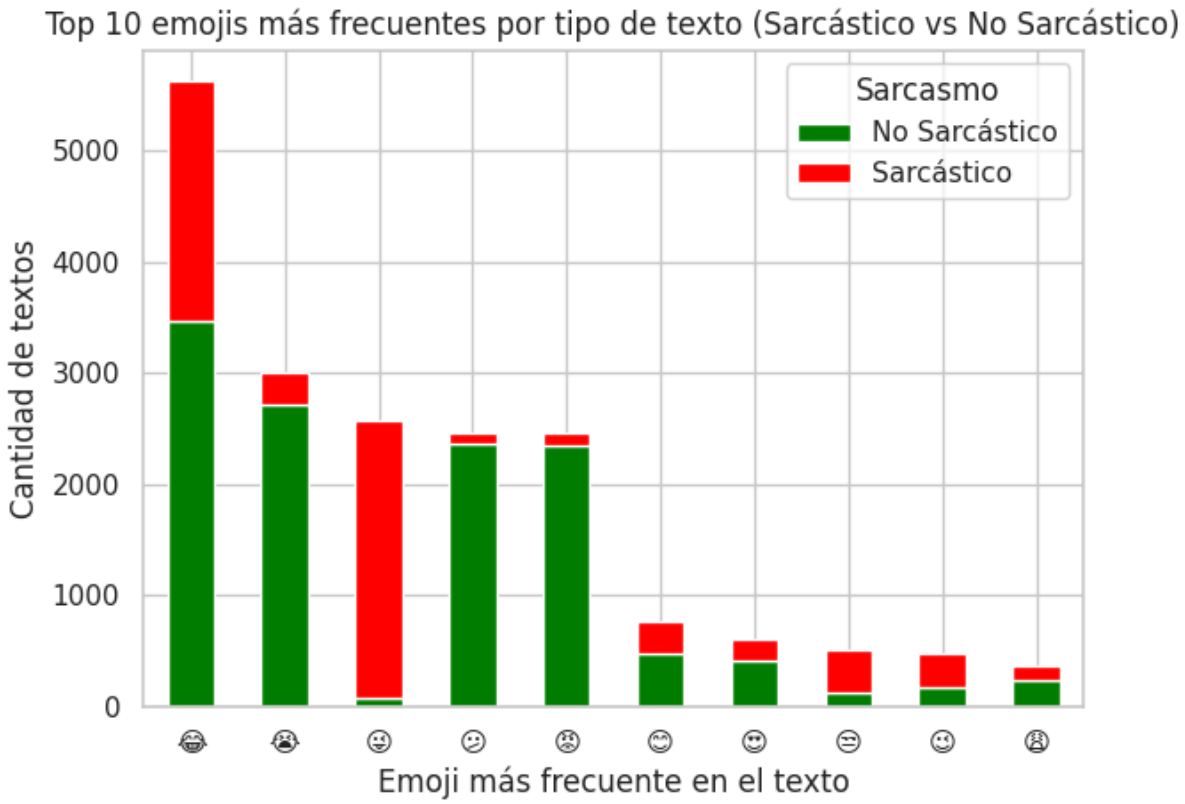
¿Son errores de carga o son reales?

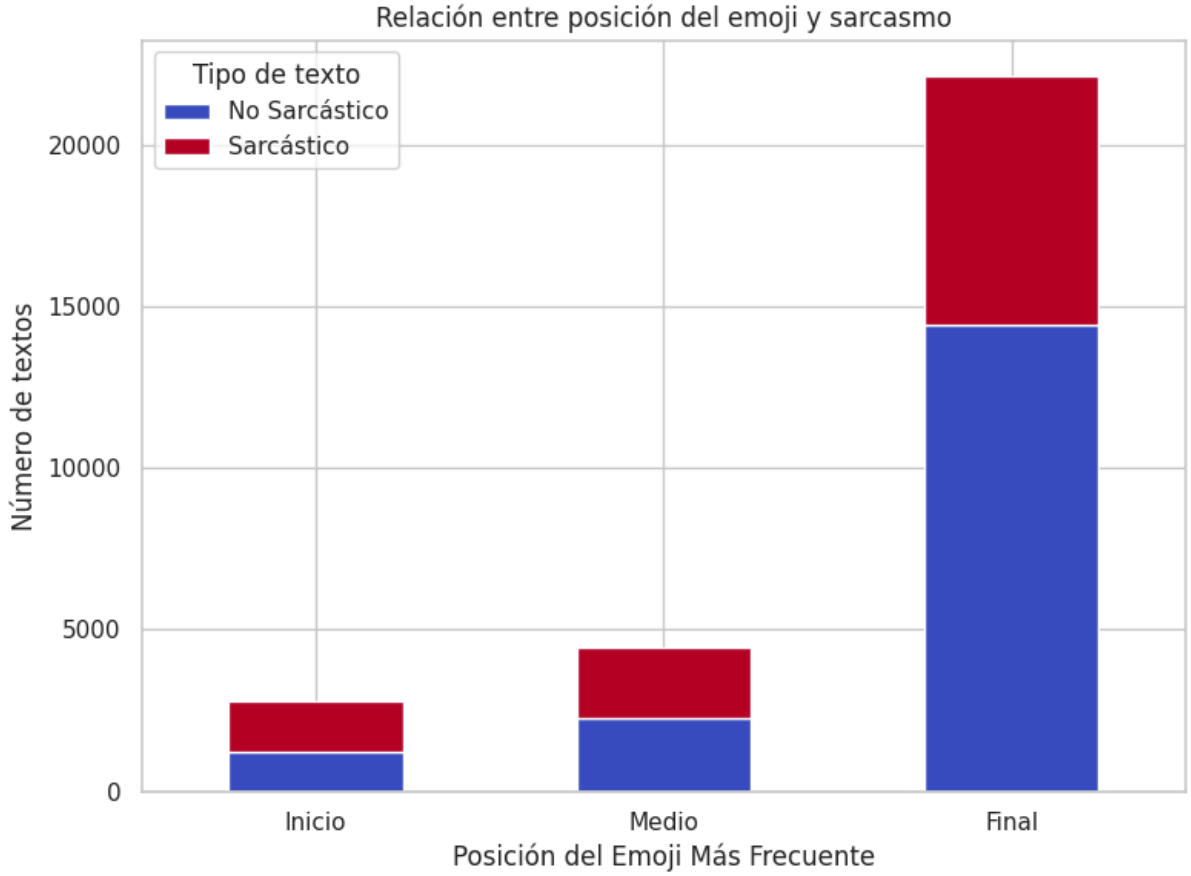
MaxEmojiNumOccurence

- En redes sociales es común repetir emojis para enfatizar una emoción o burla. No es necesariamente un error de carga. Puede ser una exageración intencional.

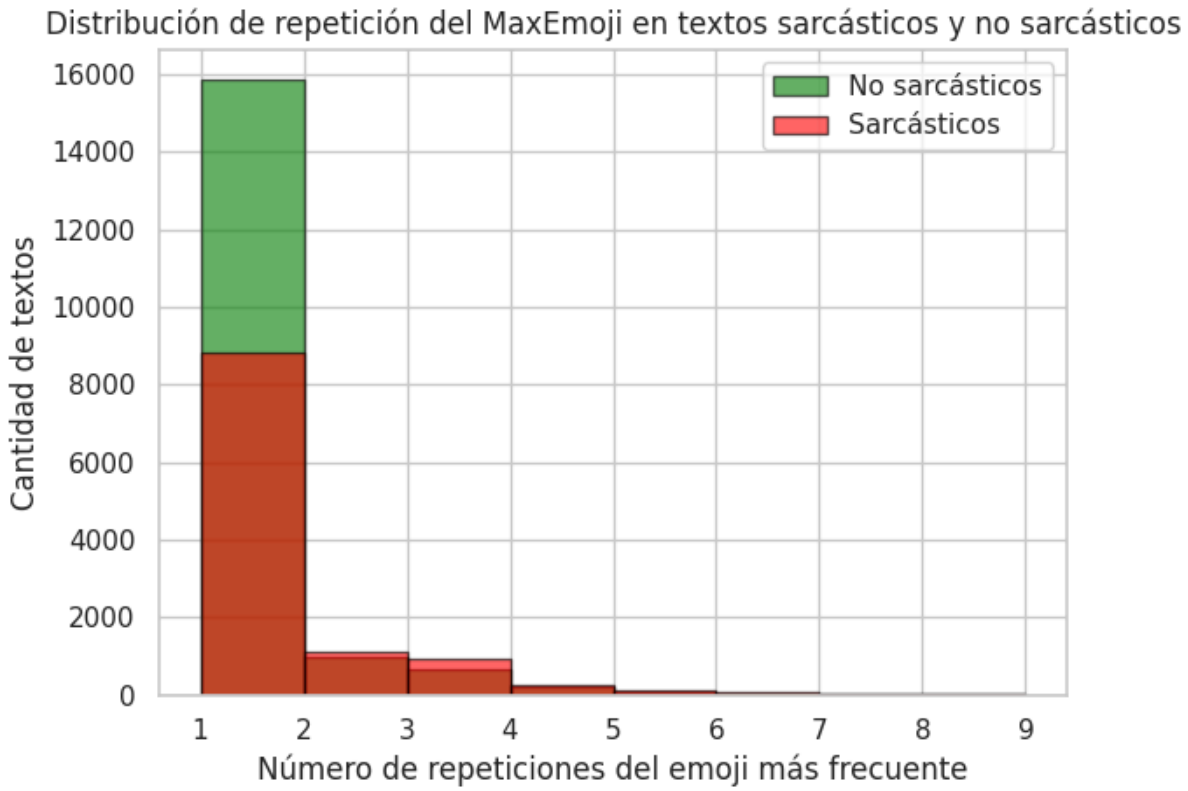
3. Visualización

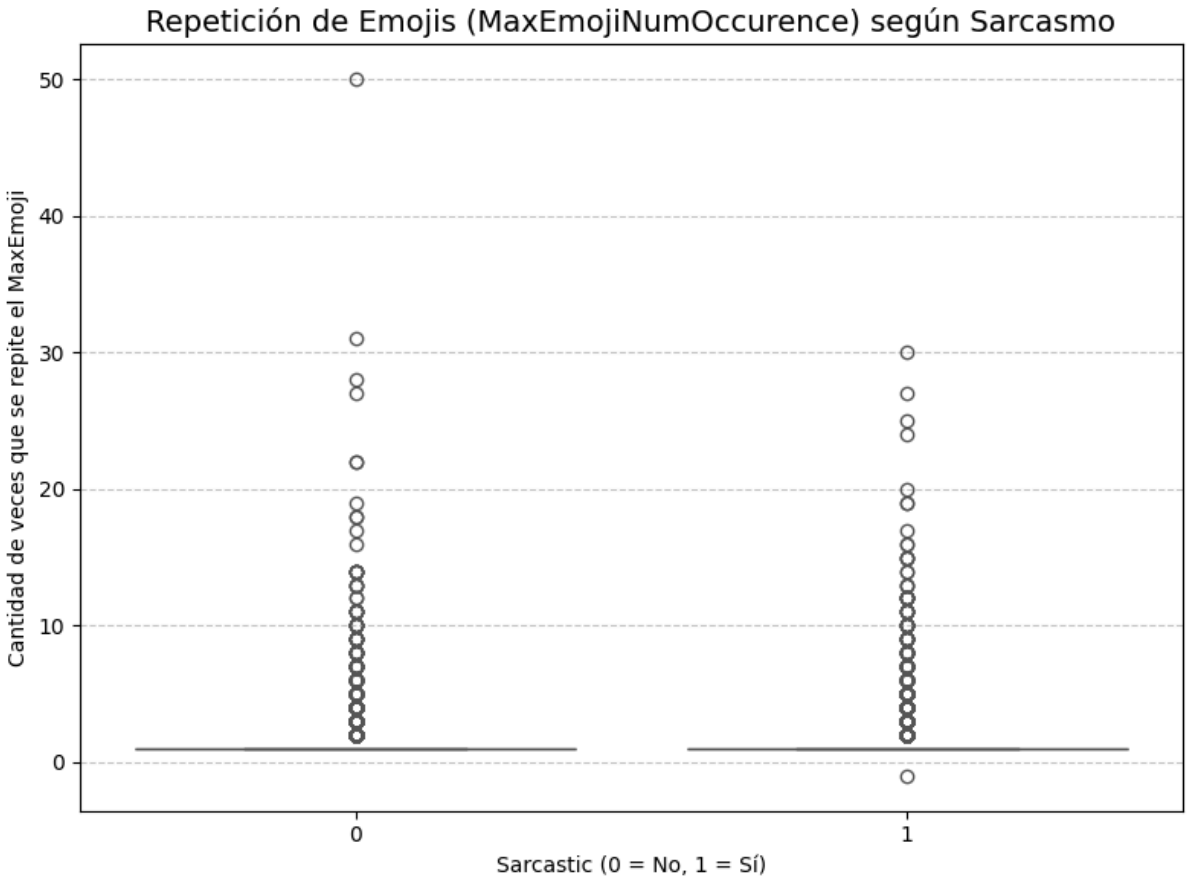
- **Variables categóricas**


Top 10 emojis más frecuentes por tipo de texto (Sarcástico vs No Sarcástico)
Sarcasmo
No Sarcástico
Sarcástico
Cantidad de textos
0
1000
2000
3000
4000
5000
Emoji más frecuente en el texto


Relación entre posición del emoji y sarcasmo
Tipo de texto
No Sarcástico
Sarcástico
Número de textos
0
5000
10000
15000
20000
Inicio
Medio
Final
Posición del Emoji Más Frecuente

- **Variables numéricas**



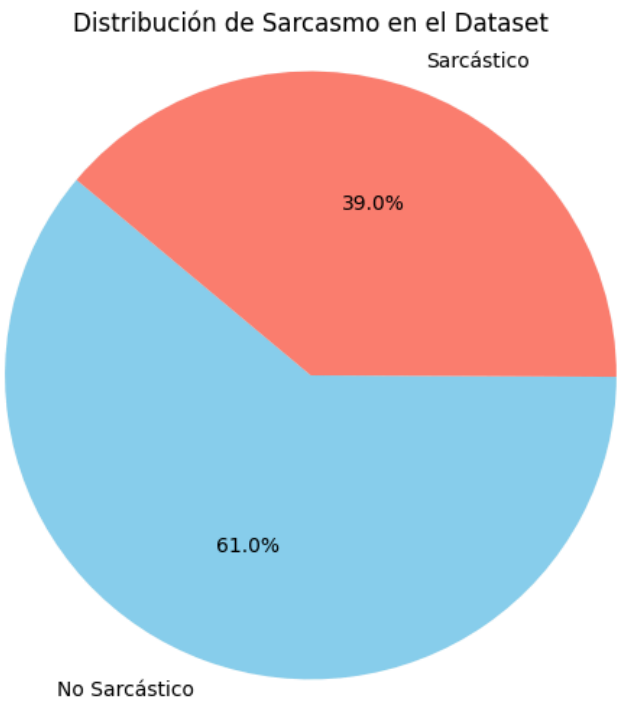


4. Encuentra un problema potencial en tus dato

Balanceo entre etiquetas

- El conjunto de datos trata de un problema de clasificación binaria supervisada, por lo que es necesario tener muestras equilibradas para cada clase.

- El conjunto de datos no está completamente balanceado. Hay un desequilibrio moderado, con más textos no sarcásticos que sarcásticos.



Características importantes

Característica	**Relevancia**
Text	Texto base: fundamental para NLP.
MaxEmoji	Refleja la emoción dominante.
MaxEmojiNumOccurence	Indica intensidad emocional (repetición sarcástica).
MaxEmojiPos	Posición del emoji: contexto del sarcasmo.
TextVader, TextSWN, TextTextBlob	Sentimiento del texto: posible incongruencia
EmojiVader, EmojiSWN, EmojiTextBlob, ESR	Sentimiento de los emojis: contraste con el texto.
Sarcastic	Etiquetado

Características descartables

Característica	**Relevancia**

Emojis	Si se usa MaxEmoji y su análisis, esta columna completa puede ser redundante.
MEmojiWN, MEVader, METB	Capturan el mismo fenómeno (Análisis de sentimientos) con distintos métodos.

Relación con el tiempo y variación con el tiempo

No. Este no es un problema de series temporales, el conjunto de datos no contiene campos como timestamp, fecha, hora, etc. Podría entrar dentro de está categoría si el conjunto de datos tuviera datos cómo la fecha y hora en la que fue publicada el post o comentario. Sin embargo, este no es el caso.

Visión Artificial

Hay miles de muestras por clase, 17929 registros de datos etiquetados cómo no sarcásticos y 11448 cómo sarcásticos, lo cual es suficiente para generalizar un modelo de machine learning si se representa adecuadamente.

Calidad de los datos

- **Outliers**: Algunas columnas tienen valores extremos (emojis repetidos 50 veces, puntajes de sentimiento mayores a 10).
- **Moderado desbalance de clases**: 61% no sarcástico, 39% sarcástico. Esto puede afectar a los modelos si no se ajustan pesos o se balancea la clase durante el entrenamiento.
- **Redundancia entre atributos**: Hay varias columnas que capturan el mismo fenómeno con distintos métodos (MEmojiWN, MEVader, METB, ESR).

5. Conclusión

- ### El sarcasmo es un fenómeno lingüístico complejo

El principal hallazgo es que el sarcasmo no se correlaciona fuertemente con métricas simples de sentimiento, ni del texto ni de los emojis. Esto refuerza la idea de que:

- El sarcasmo no puede reducirse a "emociones negativas".

- Puede estar disfrazado como positivo, neutral o incluso entusiasta.
- Requiere comprender contradicciones semánticas y contexto para ser detectado.

Detectar sarcasmo necesita enfoques más profundos que análisis de polaridad; se requiere modelado contextual o pragmático del lenguaje.

- **Los emojis no siempre refuerzan el tono del texto**

  Las emociones expresadas por los emojis no están alineadas con las del texto. De hecho:

  - A veces amplifican el sarcasmo al generar un contraste con lo que se dice.
  - Otras veces se usan de forma irónica, desincronizada del contenido literal.

  Los emojis deben analizarse como entidades semióticas independientes, no solo como reforzadores del texto.

- **Existen variables redundantes o derivadas del mismo proceso**

  Se observan fuertes correlaciones entre:

  - Sentimientos agregados y máximos de emojis.
  - Múltiples métodos de análisis (Vader, TextBlob, SWN) aplicados a las mismas unidades.

  Es importante realizar selección de variables y reducir dimensionalidad antes de aplicar modelos, para evitar sobreajuste o multicolinealidad.