TRANSFORMERS APLICADO A LA DETECCIÓN Y ANÁLISIS DE AGRESIVIDAD EN REDES SOCIALES

Victor Manuel Gómez Espinosa

Asesores:

Dr. Victor Muñiz Sánchez

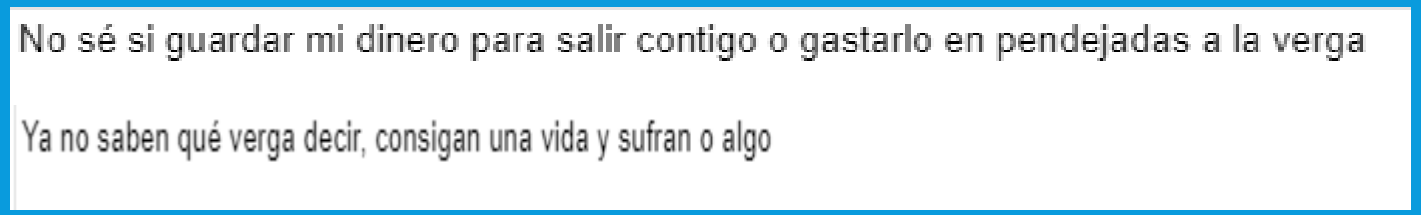
Dr. Adrián Pastor López Monroy

INTRODUCCIÓN

- **Identificación de agresividad en tweets en español de México (MEX-A3T 2020)**

INTRODUCCIÓN

- **Identificar la agresividad no es un problema fácil**, puesto que no sólo depende de la presencia o ausencia de palabras.



Ejemplos del corpus MEX-A3T 2020: arriba) tweets no agresivo. abajo) tweet agresivo

DATOS

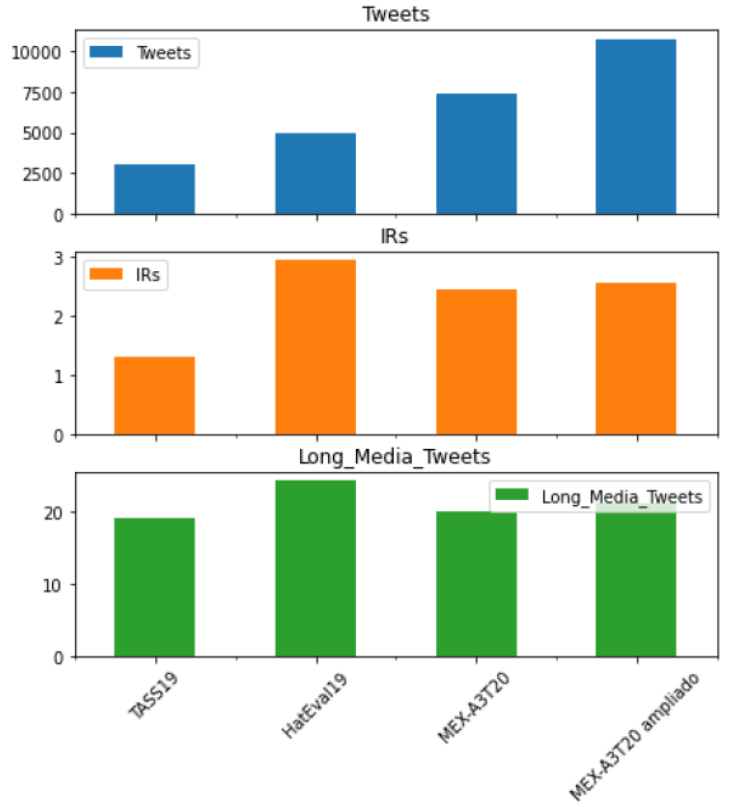


Figura 3.1.1: Comparación de los conjuntos de datos empleados por cantidad de tweets, tasa de desbalance, longitud media de los tweets.

DATOS

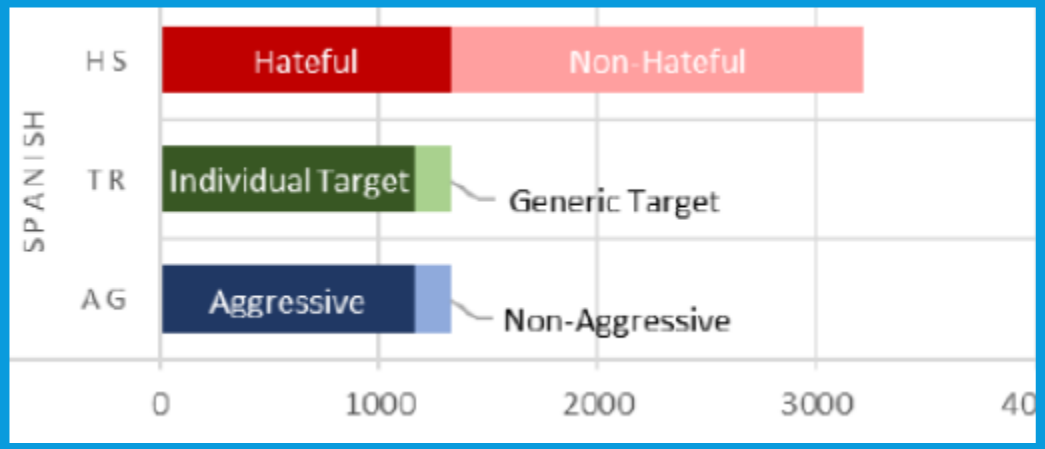
En el caso de odio, lo definen como una expresión en términos de estereotipos irrespetuosos o de misoginia más que el uso de un lenguaje vulgar, por lo cual no necesariamente está relacionado con la agresividad (Figura 3.1.5).

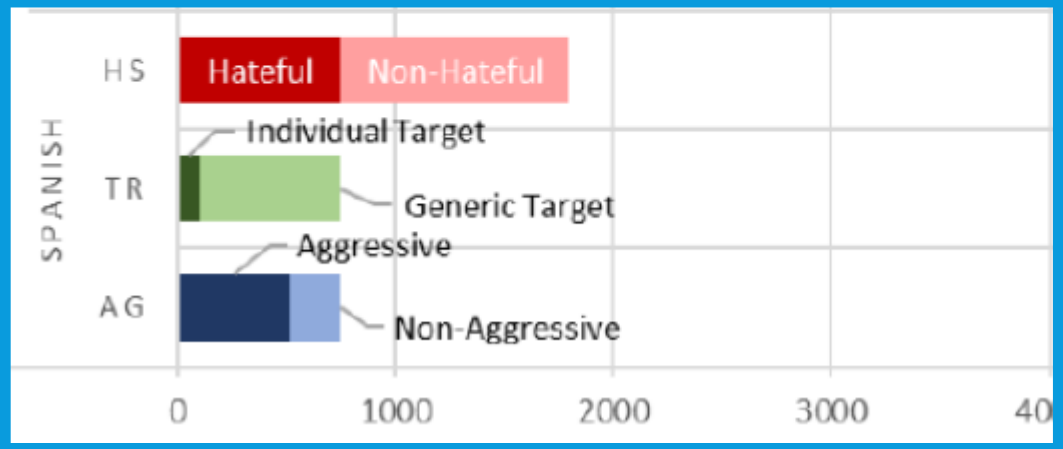
```
Padres indocumentados realizaron una huelga de hambre para pedir ser liberados https://t.co/ZPuBt8kUrK https://t.co/zRentrV0Kq'

'@MaivePerez Llora te lo.mereces por zorra'
```

Figura 3.1.5: Ejemplos del corpus **HatEval19**: Arriba, tweet no agresivo, dirigido a grupos, sin odio. Abajo, tweet agresivo, dirigido a una mujer, con odio.

DATOS





Distribución de las categorías en español de tweets en el HatEval19 dirigidos a: a) una mujer, b) a grupos de personas

PRE-PROCESO

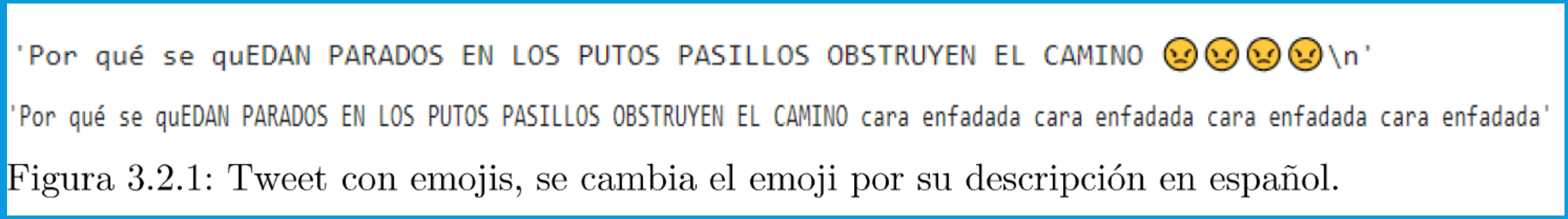


Figura 3.2.1: Tweet con emojis, se cambia el emoji por su descripción en español.

CARACTERÍSTICAS

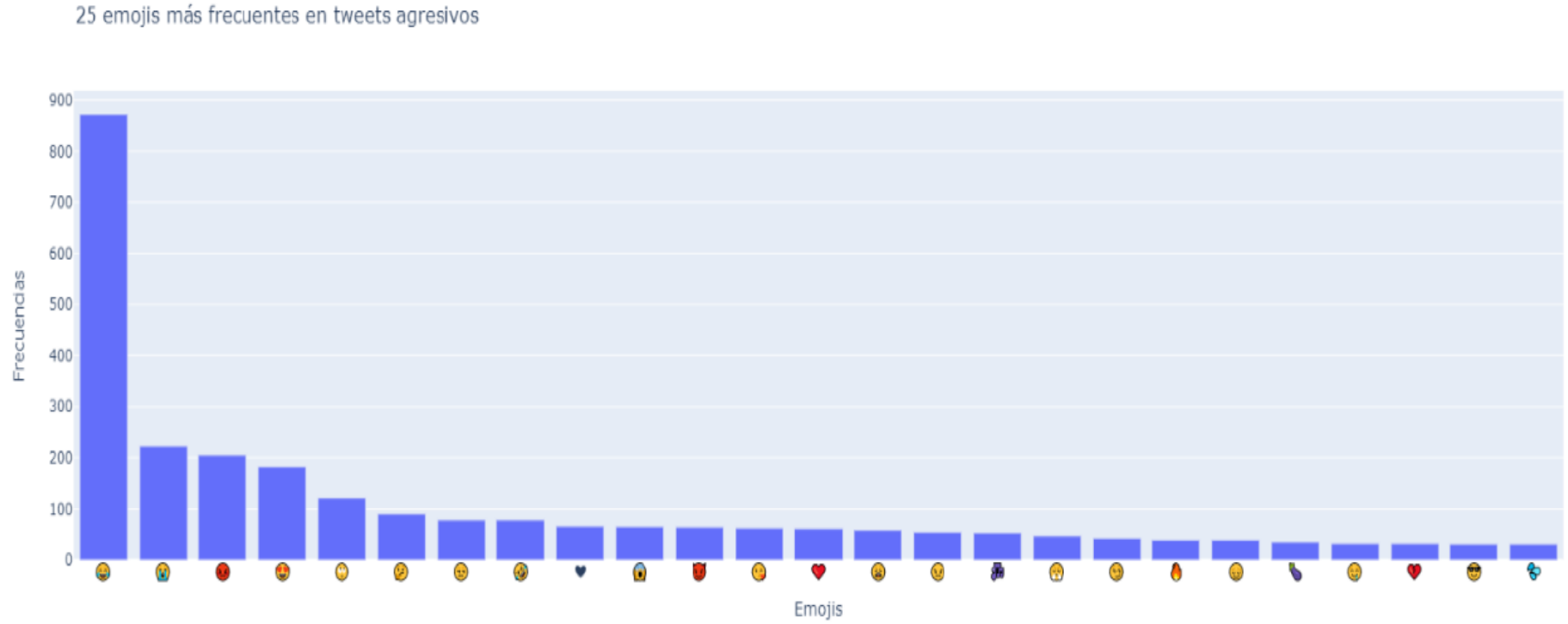


Figura 3.3.6: 25 emojis más frecuentes en tweets agresivos en el **MEX-A3T20 ampliado**.

CARACTERÍSTICAS

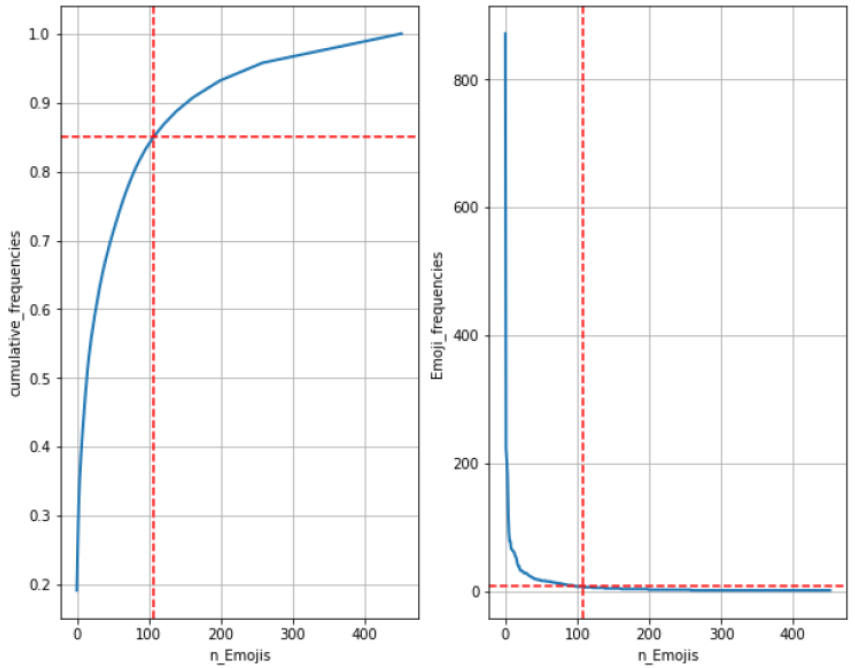


Figura 3.3.5: Selección de número de emojis como características, equivalente al 85% de los emojis más frecuentes en la categoría de agresividad en el **MEX-A3T20 ampliado**.

CARACTERÍSTICAS

$$EmojiNeg = (1 - Lexicon\_score)/2 \quad (1)$$

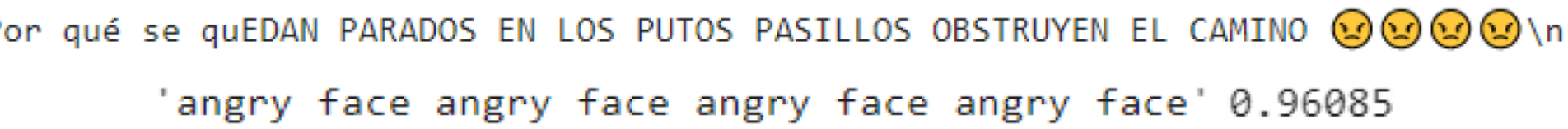


Figura 3.3.7: Tweet con emojis, se obtienen las descripciones de los emojis en inglés y se utiliza la librería Vader en Python (Hutto 2014) para obtener la probabilidad de que sea negativo.

CARACTERÍSTICAS

'Por qué se quEDAN PARADOS EN LOS PUTOS PASILLOS OBSTRUYEN EL CAMINO 😠😠😠😠\n'

😠	4.000000
EmojiNeg	0.960850
Women	0.001665
Hate	0.004554
Negative	0.997410
Aggressiveness	0.000726

Figura 3.3.1: Características de un tweet no agresivo del corpus **MEX-A3T20 ampliado**.

EXPERIMENTOS Y RESULTADOS

Modelo	***F1***
BERT_Emojis	79.2516
BERT_3Features	80.1920
Ganador_MEX-A3T20	**80.2937**
XGB_4Features	80.4400
Ensamble_XGB+BERT	80.4597
XGB_107Emojis_5Features	80.6888
MEX-A3T20_ampliado	81.06151

Tabla 3.5.1: Resultados de experimentos realizados según el score F1.

EXPERIMENTOS Y RESULTADOS

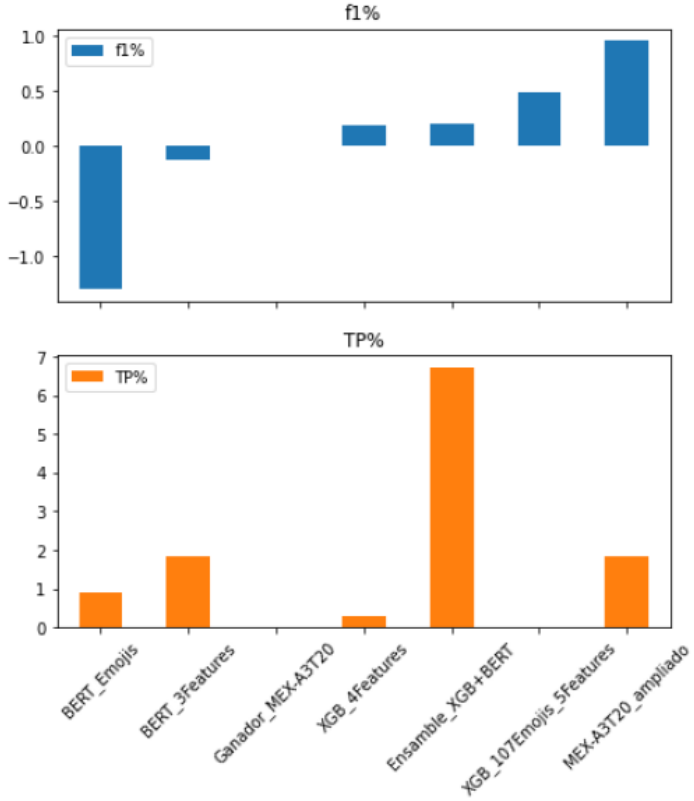


Figura 3.5.1: Experimentos realizados: parte superior, cambio porcentual respecto al score f1. Parte inferior, cambio porcentual respecto a los verdaderos positivos (tweets agresivos) del modelo ganador del **MEX-A3T20**.

ANÁLISIS E INTERPRETACIÓN DE LOS MODELOS

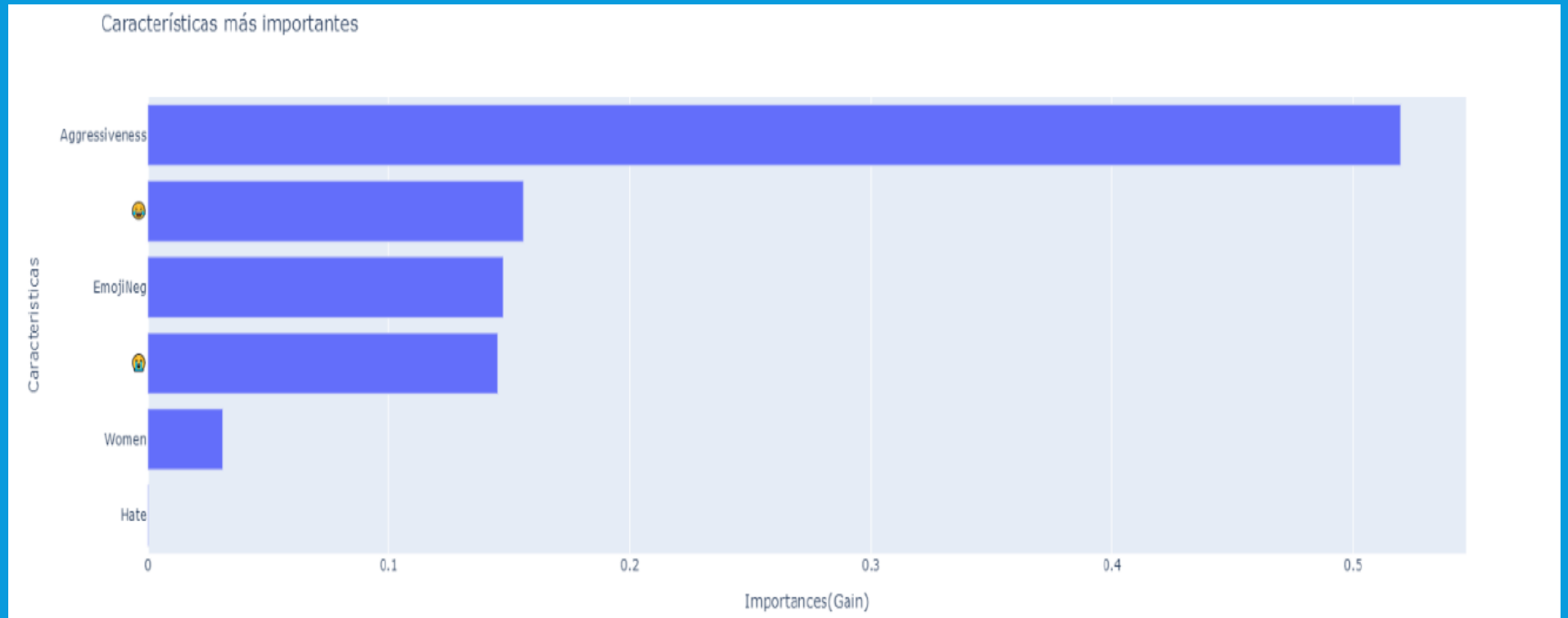


Figura 4.1: Características más importantes (según el criterio del GAIN) para el modelo XGBoost_107Emojis_5Features.

CRONOGRAMA

			Nombre de tarea	Duración	Comienz	Fin	Predecesora
1			Revisión de la literatura del estado del arte en Transformers y agresividad	55 días	mar 19/1/21	dom 14/3/21	
2			Revisión de las implementaciones de Transformers para tareas de análisis de textos basado en PyTorch	55 días	mar 19/1/21	dom 14/3/21	
3	✓		Exploración del corpus de agresividad MEX-A3T y código ganador	55 días	mar 19/1/21	dom 14/3/21	
4			Selección y adaptación de métodos en el estado del arte en Transformers para el caso específico de la tarea de detección de agresividad	14 días	lun 15/3/21	dom 28/3/21	1
5			Programación, entrenamiento y validación del método seleccionado	35 días	lun 29/3/21	dom 2/5/21	4
6			Evaluación de resultados	14 días	lun 3/5/21	dom 16/5/21	5
7			Escritura de tesis	107 días	jue	vie