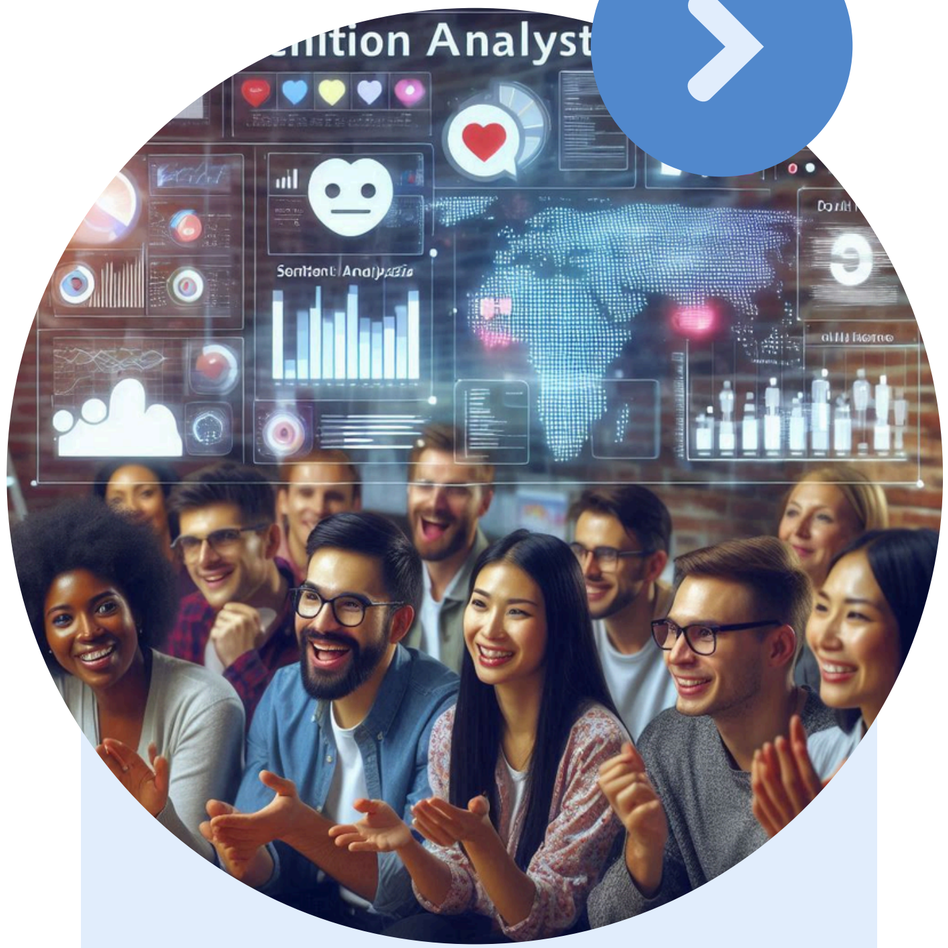
ML CLASSIFICATION TWEETS

BOOTCAMP XPERIENCE

GRUPO 2

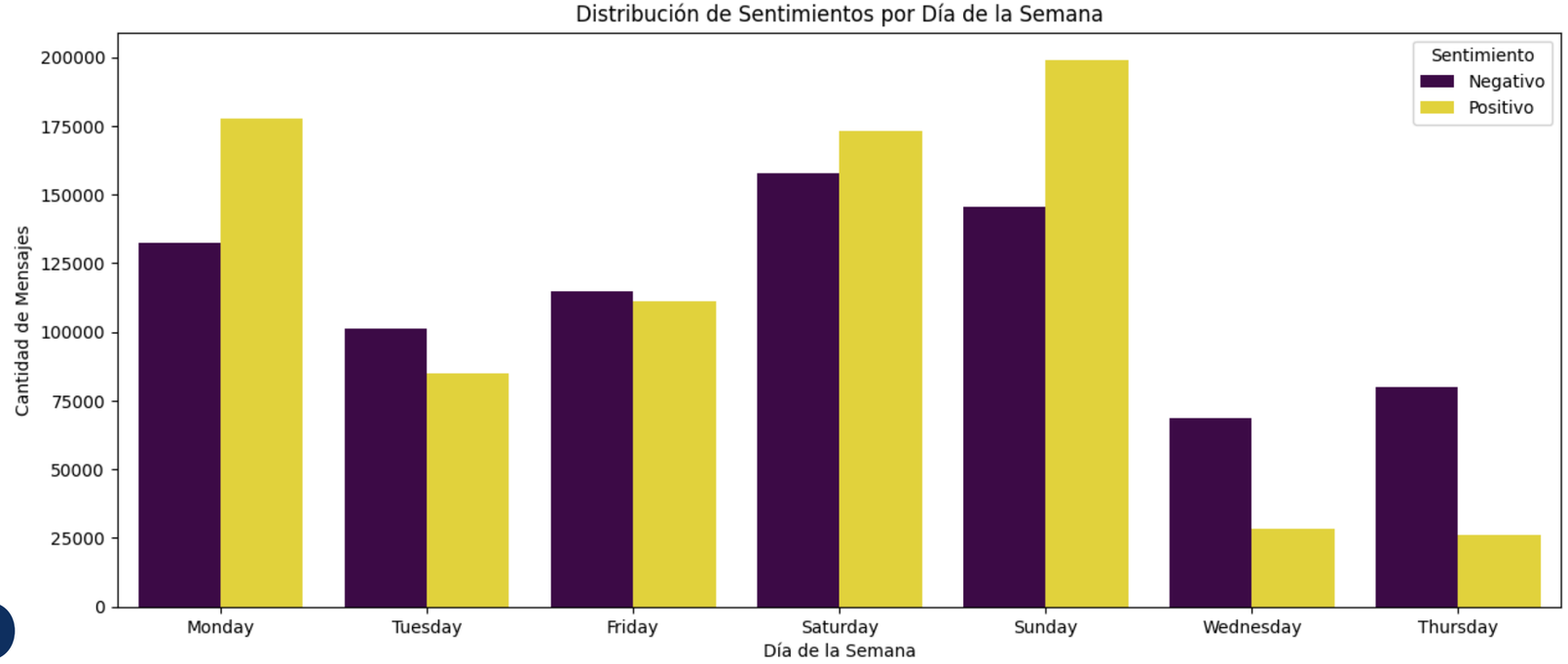


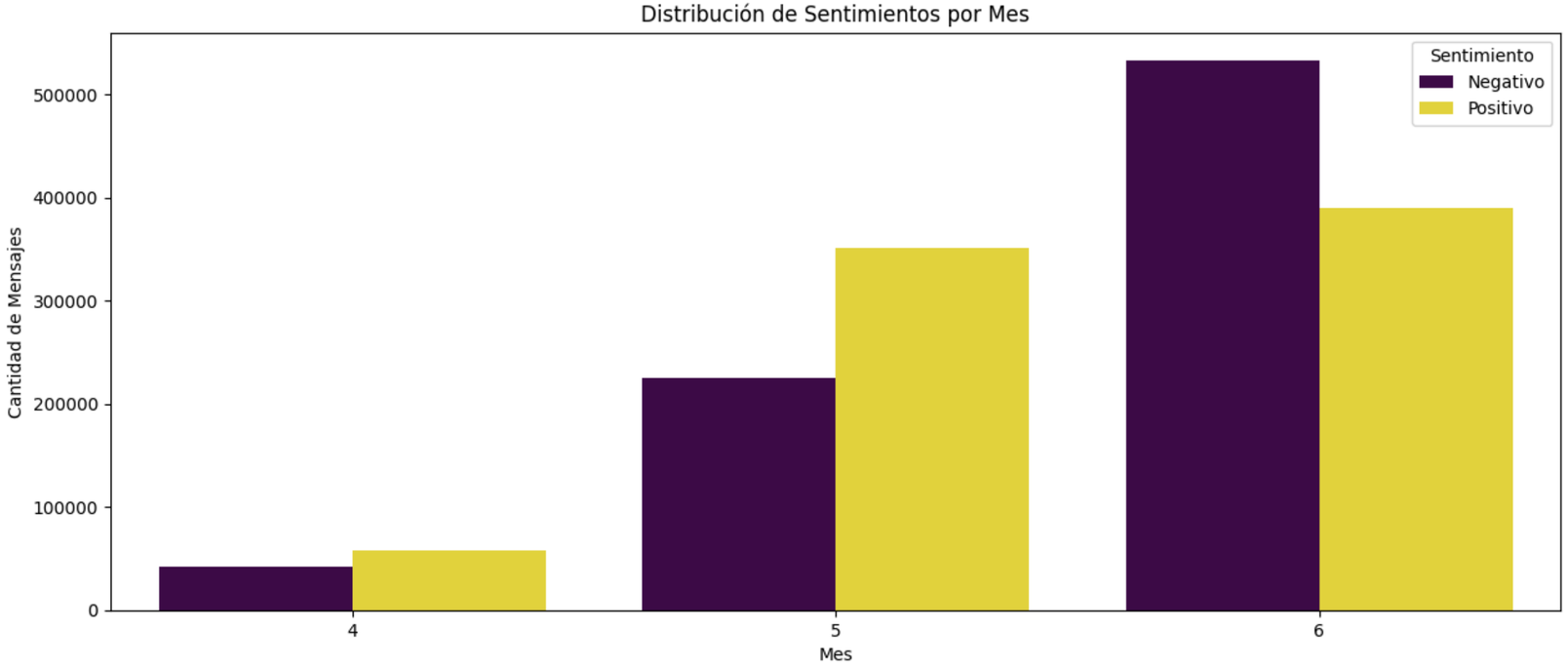
Introduction

Este proyecto analiza los sentimientos en tweets, clasificándolos como positivos o negativos. Usamos un conjunto de 1,600,000 tweets preprocesados, aplicando diversas técnicas para limpiar el texto y extraer características clave. Luego, entrenamos modelos de clasificación para predecir el sentimiento positivo o negativo de los nuevos tweets que se presenten.

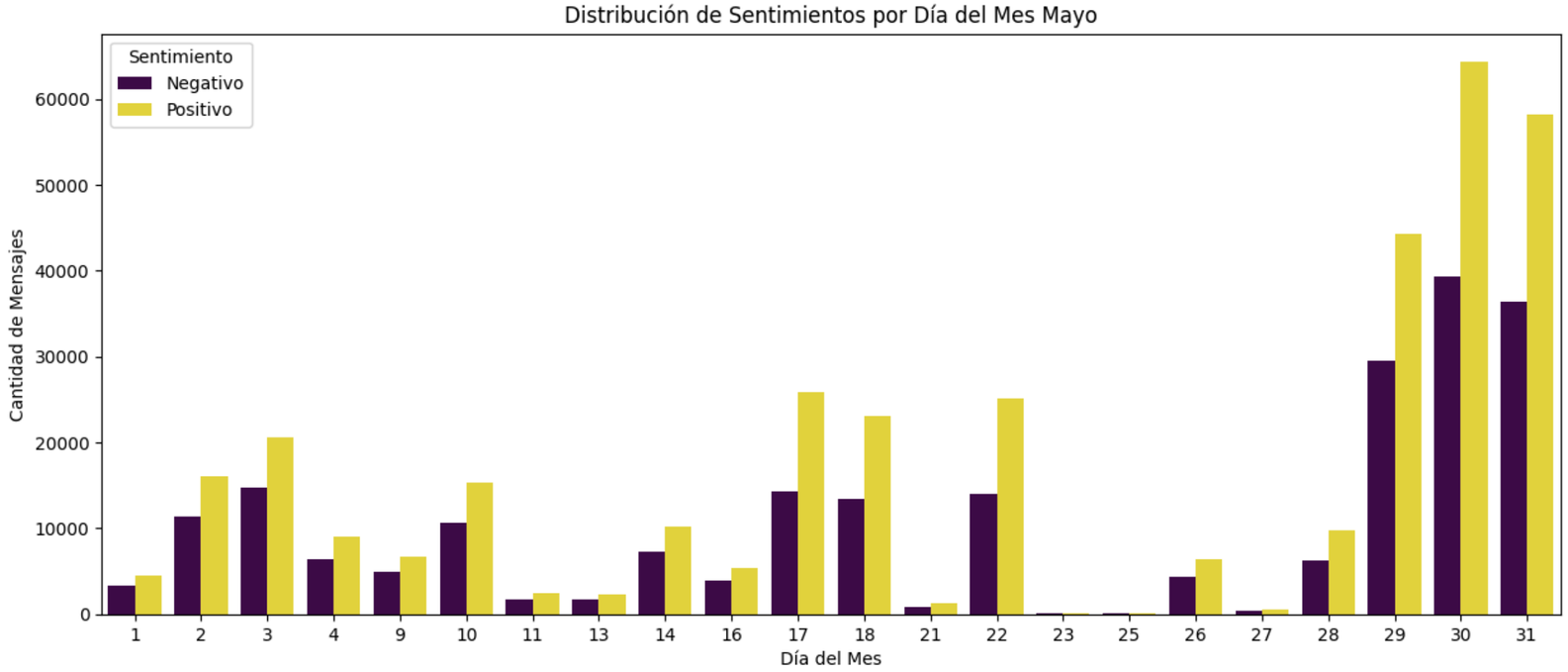
El modelo se despliega con Streamlit, ofreciendo una interfaz interactiva para visualizar los resultados en tiempo real.

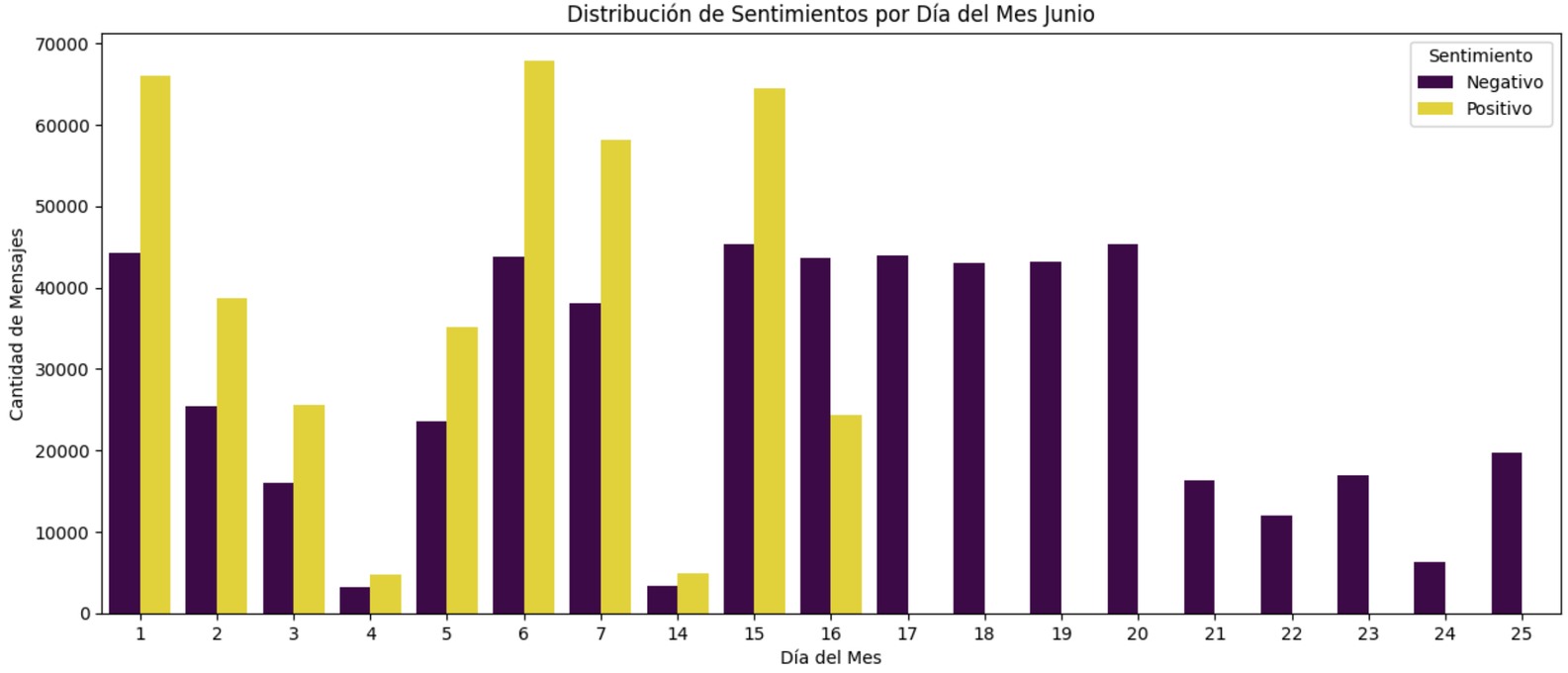
ANÁLISIS EXPLORATORIO INICIAL DEL DATASET (Por Fechas)



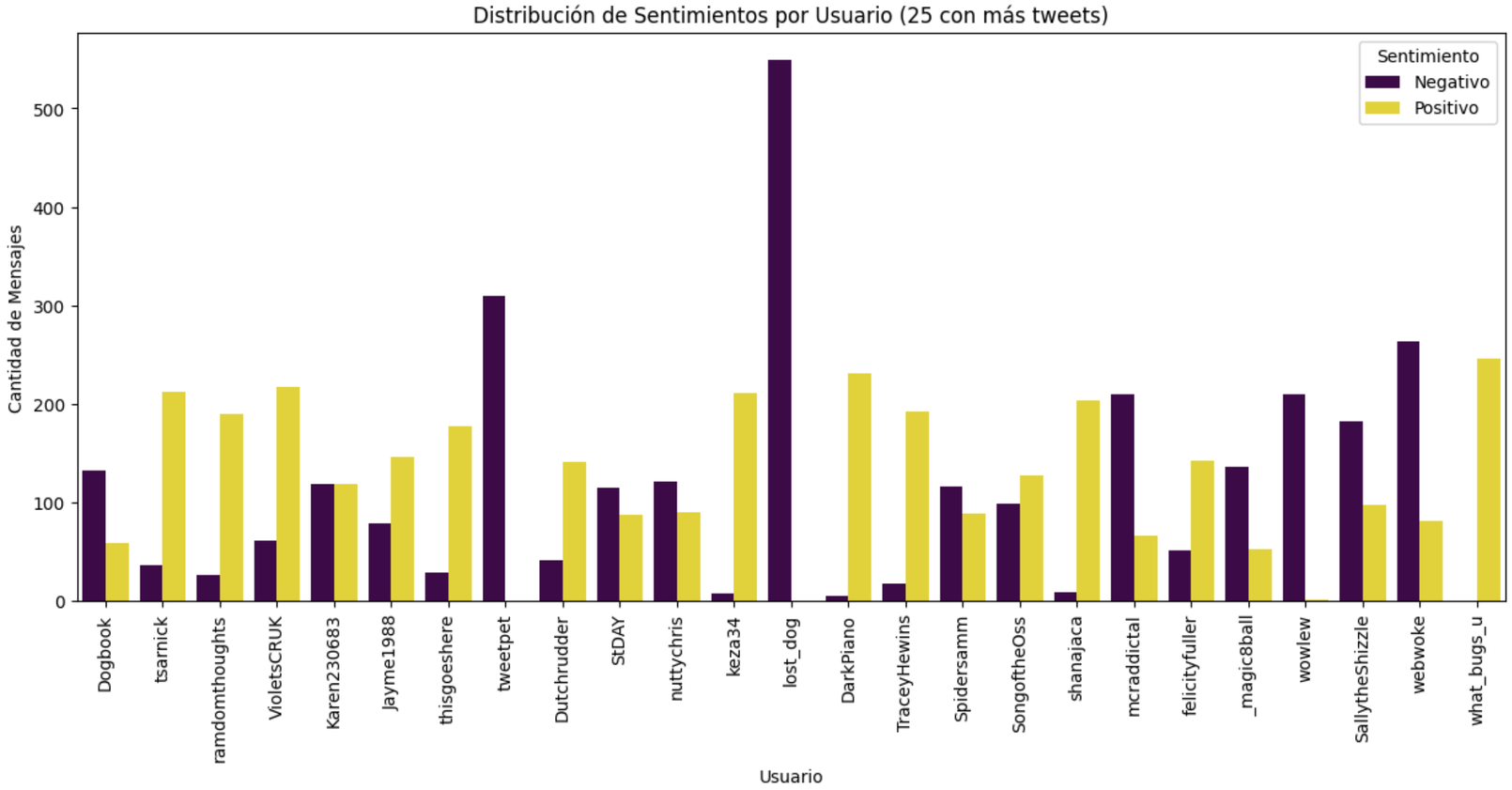


ANÁLISIS EXPLORATORIO INICIAL DEL DATASET (Por Fechas)





ANÁLISIS EXPLORATORIO INICIAL DEL DATASET (Por Usuarios)



	target	ids	date	flag	user	text	month	day	day_name
43934	0	1676704158	Fri May 01 22:54:02 PDT 2009	NO_QUERY	lost_dog	@NyleW I am lost. Please help me find a good h...	5	1	Friday
45573	0	1677189389	Sat May 02 00:51:35 PDT 2009	NO_QUERY	lost_dog	@SallyD I am lost. Please help me find a good ...	5	2	Saturday
46918	0	1677519173	Sat May 02 02:30:50 PDT 2009	NO_QUERY	lost_dog	@zuppaholic I am lost. Please help me find a g...	5	2	Saturday
47948	0	1677752995	Sat May 02 03:47:51 PDT 2009	NO_QUERY	lost_dog	@LOSTPETUSA I am lost. Please help me find a g...	5	2	Saturday
50571	0	1678544903	Sat May 02 07:02:28 PDT 2009	NO_QUERY	lost_dog	@JeanLevertHood I am lost. Please help me find...	5	2	Saturday
...	...	...	...	...	...	...	...	...	...
792408	0	2326272045	Thu Jun 25 06:48:18 PDT 2009	NO_QUERY	lost_dog	@trooppetrie I am lost. Please help me find a ...	6	25	Thursday
793313	0	2326588770	Thu Jun 25 07:14:42 PDT 2009	NO_QUERY	lost_dog	@Carly_FTS I am lost. Please help me find a go...	6	25	Thursday
793609	0	2326689658	Thu Jun 25 07:22:51 PDT 2009	NO_QUERY	lost_dog	@inathlone I am lost. Please help me find a go...	6	25	Thursday
798607	0	2328636087	Thu Jun 25 09:49:04 PDT 2009	NO_QUERY	lost_dog	@Kram I am lost. Please help me find a good ho...	6	25	Thursday
799404	0	2328965183	Thu Jun 25 10:11:34 PDT 2009	NO_QUERY	lost_dog	@W_Hancock I am lost. Please help me find a go...	6	25	Thursday

549 rows × 9 columns

ANÁLISIS EXPLORATORIO INICIAL DEL DATASET (Por Usuarios)

	target	ids	date	flag	user	text	month	day	day_name
363302	0	2047801265	Fri Jun 05 14:12:05 PDT 2009	NO_QUERY	webwoke	auchh, drop by 1 (32)elitestv.com	6	5	Friday
366528	0	2048883634	Fri Jun 05 15:56:07 PDT 2009	NO_QUERY	webwoke	auchh, drop by 1 (7)pedeee.com	6	5	Friday
366529	0	2048883882	Fri Jun 05 15:56:08 PDT 2009	NO_QUERY	webwoke	auchh, drop by 1 (17)rumahabi.com	6	5	Friday
366596	0	2048903368	Fri Jun 05 15:58:07 PDT 2009	NO_QUERY	webwoke	auchh, drop by 1 (18)twitter.com	6	5	Friday
366598	0	2048903446	Fri Jun 05 15:58:07 PDT 2009	NO_QUERY	webwoke	auchh, drop by 1 (19)yehia.org	6	5	Friday
...	...	...	...	...	...	...	...	...	...
1505837	4	2072460652	Sun Jun 07 21:04:09 PDT 2009	NO_QUERY	webwoke	ohh yesss move up by 3 99. mybloglog.com	6	7	Sunday
1505840	4	2072460726	Sun Jun 07 21:04:09 PDT 2009	NO_QUERY	webwoke	Gooo... move up by 3 100. digg.com	6	7	Sunday
1506013	4	2072494733	Sun Jun 07 21:08:09 PDT 2009	NO_QUERY	webwoke	GoGoGo... move up by 2 105. ardhindie.com	6	7	Sunday
1506014	4	2072494818	Sun Jun 07 21:08:09 PDT 2009	NO_QUERY	webwoke	uhuiii... move up by 2 106. seo-guy.com	6	7	Sunday
1506018	4	2072494917	Sun Jun 07 21:08:10 PDT 2009	NO_QUERY	webwoke	ilovegoogle, move up by 2 107. wordpress.com	6	7	Sunday

345 rows × 9 columns

	target	ids	date	flag	user	text	month	day	day_name
9559	0	1548797240	Fri Apr 17 22:00:01 PDT 2009	NO_QUERY	tweetpet	@tweetchild Clean Me!	4	17	Friday
9560	0	1548797247	Fri Apr 17 22:00:01 PDT 2009	NO_QUERY	tweetpet	@tweetchild Clean Me!	4	17	Friday
9561	0	1548797393	Fri Apr 17 22:00:02 PDT 2009	NO_QUERY	tweetpet	@chromachris Clean Me!	4	17	Friday
9563	0	1548797501	Fri Apr 17 22:00:03 PDT 2009	NO_QUERY	tweetpet	@reatlas Clean Me!	4	17	Friday
9564	0	1548797565	Fri Apr 17 22:00:03 PDT 2009	NO_QUERY	tweetpet	@chromachris Clean Me!	4	17	Friday
...	...	...	...	...	...	...	...	...	...
49678	0	1678233051	Sat May 02 06:00:37 PDT 2009	NO_QUERY	tweetpet	@amateurdelta54 Clean Me!	5	2	Saturday
49679	0	1678233110	Sat May 02 06:00:38 PDT 2009	NO_QUERY	tweetpet	@littleblue62 Clean Me!	5	2	Saturday
49680	0	1678233215	Sat May 02 06:00:39 PDT 2009	NO_QUERY	tweetpet	@Shawn1976 Clean Me!	5	2	Saturday
49693	0	1678235166	Sat May 02 06:01:02 PDT 2009	NO_QUERY	tweetpet	@TKgFMb hungry. Type 'feed' to feed me...	5	2	Saturday
49694	0	1678235231	Sat May 02 06:01:03 PDT 2009	NO_QUERY	tweetpet	@palfour89 hungry. Type 'feed' to feed me...	5	2	Saturday

310 rows × 9 columns

ANÁLISIS EXPLORATORIO INICIAL DEL DATASET (Por Usuarios)

- Se identificó que los cuatro usuarios con mayor actividad tienden a repetir mensajes, lo que sugiere que podrían ser bots.

- Los demás usuarios, que siguen a estos cuatro principales, presentan contenido más diverso, tanto en tweets positivos como negativos.

- Limitar la cantidad de tweets por usuario ayudaría a reducir sesgos en el modelo, mejorando su capacidad de generalización.

TRATAMIENTO DE DATOS

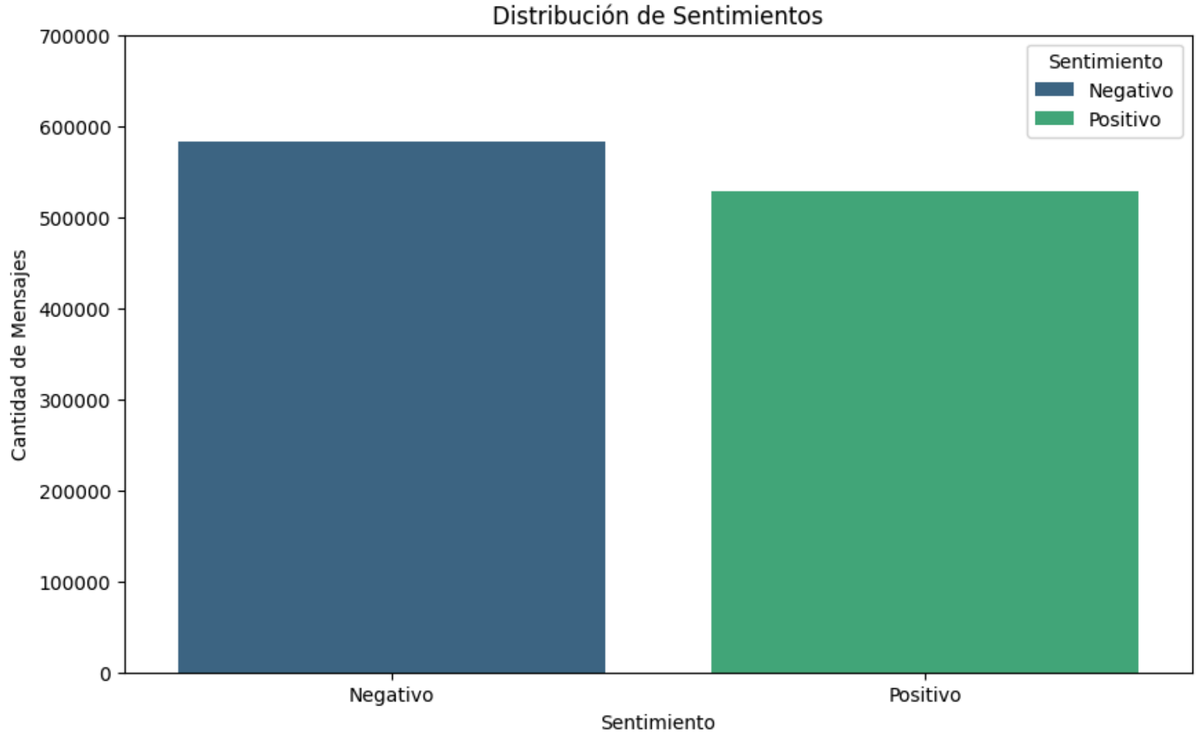
Se realizan varios tratamientos con la finalidad de limpiar nuestros comentarios y poder incluirlos en nuestro modelo de machine learning.

- Se aborda el tratamiento de hashtags, URLs, signos de puntuación, emoticones, menciones, apóstrofes, mayúsculas, entre otros.

- Eliminación de stop words: Se abordó la eliminación de palabras comunes que no aportan significado relevante al análisis, como artículos, preposiciones y conjunciones.

- Eliminación de valores faltantes y duplicados, a lo largo de todo el tratamiento de los textos del dataset.

TRATAMIENTO DE DATOS



	target	text	clean_text	without_stopwords
0	0	is upset that he can't update his Facebook by ...	is upset that he cant update his facebook by t...	upset cant update facebook texting might cry r...
1	0	@Kenichan I dived many times for the ball. Man...	i dived many times for the ball managed to sav...	dived many times ball managed save percent res...
2	0	my whole body feels itchy and like its on fire	my whole body feels itchy and like its on fire	whole body feels itchy like fire
3	0	@nationwideclass no, it's not behaving at all....	no its not behaving at all im mad why am i her...	behaving im mad cant see
4	0	@Kwesidei not the whole crew	not the whole crew	whole crew
...	...	...	...	...
1137238	4	ReCoVeRiNg FrOm ThE lOnG wEeKeNd	recovering from the long weekend	recovering long weekend
1137239	4	@Cliff_Forster Yeah, that does work better tha...	yeah that does work better than just waiting f...	yeah work better waiting end wonder time keep ...
1137240	4	Just woke up. Having no school is the best fee...	just woke up having no school is the best feel...	woke school best feeling ever
1137241	4	TheWDB.com - Very cool to hear old Walt interv...	very cool to hear old walt interviews	cool hear old walt interviews
1137242	4	Are you ready for your MoJo Makeover? Ask me f...	are you ready for your mojo makeover ask me fo...	ready mojo makeover ask details

1137243 rows × 4 columns

FEATURE ENGINEERING

En nuestro DataFrame, creamos variables basadas en diversas características del texto original, el texto limpio (sin ruido) y el texto sin stopwords. Estas nuevas variables se utilizarán en el modelo de predicción para mejorar su desempeño.

- Conteo de palabras totales: Número total de palabras en el texto.
- Conteo de palabras únicas: Cantidad de palabras distintas en el texto.
- Palabras positivas, negativas y neutras: Clasificación de palabras según su carga emocional.
- Entropía: Medida de la diversidad de información en el texto.
- Conteo de stopwords: Número de palabras vacías (artículos, preposiciones, etc.).
- Conteo de signos de interrogación: Cantidad de veces que aparece "¿" o "?".
- Conteo de signos de exclamación: Número de signos "¡" o "!".
- Conteo de signos de puntuación: Cantidad total de comas, puntos, etc.
- Densidad de palabras en mayúsculas: Proporción de palabras escritas completamente en mayúsculas.
- Palabras repetidas: Número de palabras que aparecen más de una vez en el texto.
- Sarcasmo: Detección de posibles patrones de ironía o sarcasmo.
- Polaridad: Medida de si el texto es positivo, negativo o neutral.
- Subjetividad: Indica si el texto expresa opiniones personales o hechos objetivos.

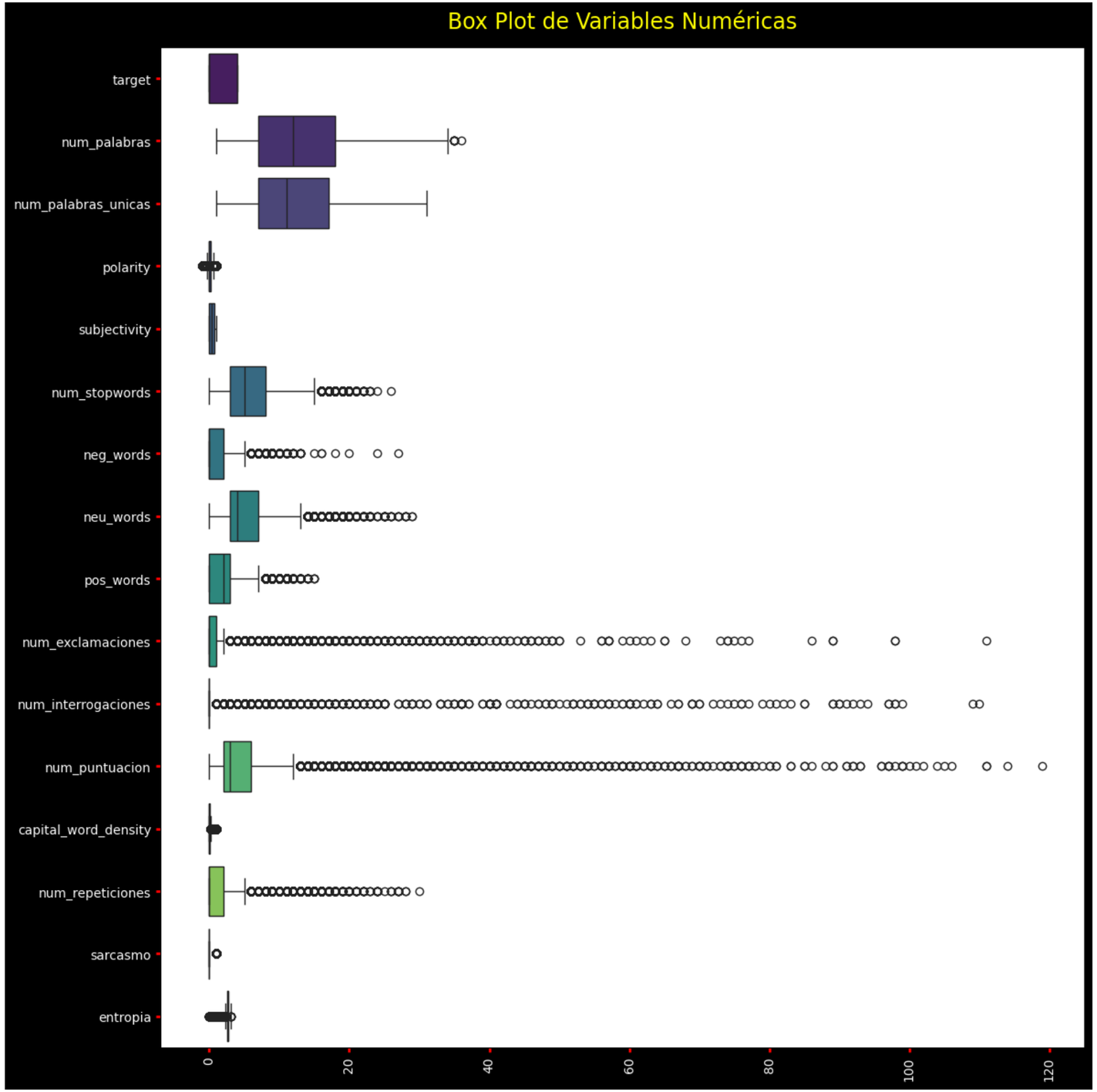
FEATURE ENGINEERING

target	text	clean_text	without_stopwords
0	is upset that he can't update his Facebook by ...	is upset that he cant update his facebook by t...	upset cant update facebook texting might cry r...
0	@Kenichan I dived many times for the ball. Man...	i dived many times for the ball managed to sav...	dived many times ball managed save percent res...
0	my whole body feels itchy and like its on fire	my whole body feels itchy and like its on fire	whole body feels itchy like fire
0	@nationwideclass no, it's not behaving at all....	no its not behaving at all im mad why am i her...	behaving im mad cant see
0	@Kwesidei not the whole crew	not the whole crew	whole crew
...	...	...	...
4	ReCoVeRiNg FrOm ThE lOnG wEeKeNd	recovering from the long weekend	recovering long weekend
4	@Cliff_Forster Yeah, that does work better tha...	yeah that does work better than just waiting f...	yeah work better waiting end wonder time keep ...
4	Just woke up. Having no school is the best fee...	just woke up having no school is the best feel...	woke school best feeling ever
4	TheWDB.com - Very cool to hear old Walt interv...	very cool to hear old walt interviews	cool hear old walt interviews
4	Are you ready for your MoJo Makeover? Ask me f...	are you ready for your mojo makeover ask me fo...	ready mojo makeover ask details

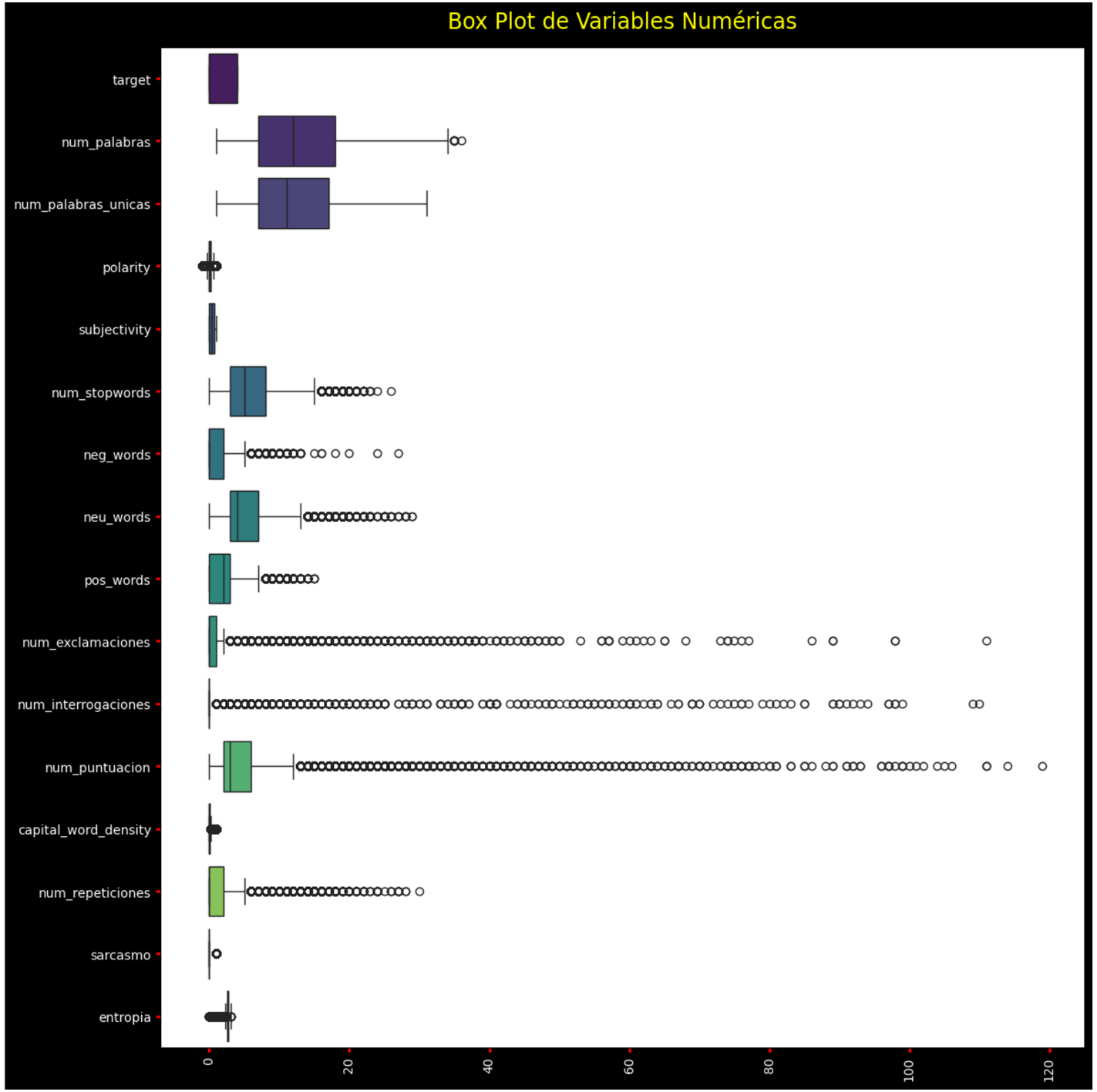
ws × 19 columns

num_palabras	num_palabras_unicas	polarity	subjectivity	num_stopwords	neg_words	neu_words	pos_words	num_exclamaciones	num_interrogaciones	num_puntuacion	capital_word_density	num_repeticiones	sarcasmo	entropia
21	21	0.000	0.000	9	5	7	0	1	0	6	0.000000	0	0	2.789718
17	16	0.500	0.500	7	0	7	3	0	0	3	0.055556	2	0	2.741038
10	10	0.200	0.400	4	3	2	2	0	0	0	0.000000	0	0	2.707497
20	18	-0.625	1.000	15	2	3	0	0	1	9	0.047619	4	0	2.689712
4	4	0.200	0.400	2	0	2	0	0	0	1	0.000000	0	0	2.216102
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	5	-0.050	0.400	2	0	3	0	0	0	0	0.000000	0	0	2.604188
26	24	0.600	0.550	16	0	5	5	0	0	4	0.074074	4	0	2.736297
11	11	1.000	0.300	6	0	2	3	0	0	1	0.000000	0	0	2.758130
7	7	0.225	0.425	2	0	3	2	1	0	9	0.000000	0	0	2.603523
11	10	0.200	0.500	6	0	3	2	0	1	1	0.000000	2	0	2.557903

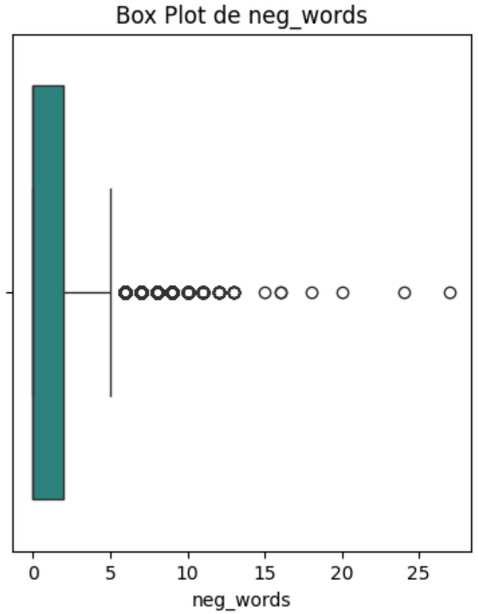
ANÁLISIS EXPLORATORIO DE DATOS (EDA)

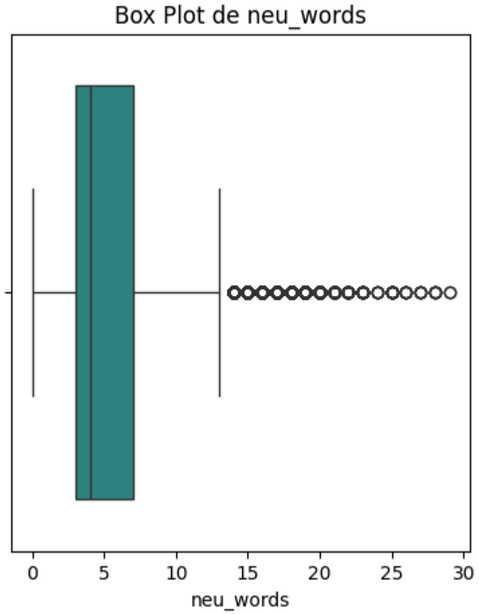


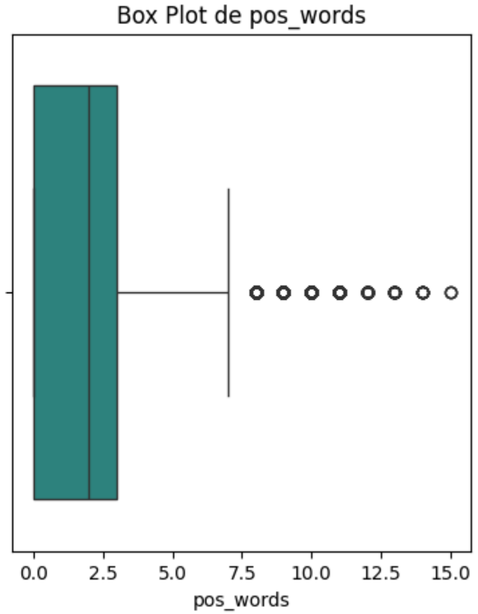
ANÁLISIS EXPLORATORIO DE DATOS (EDA)

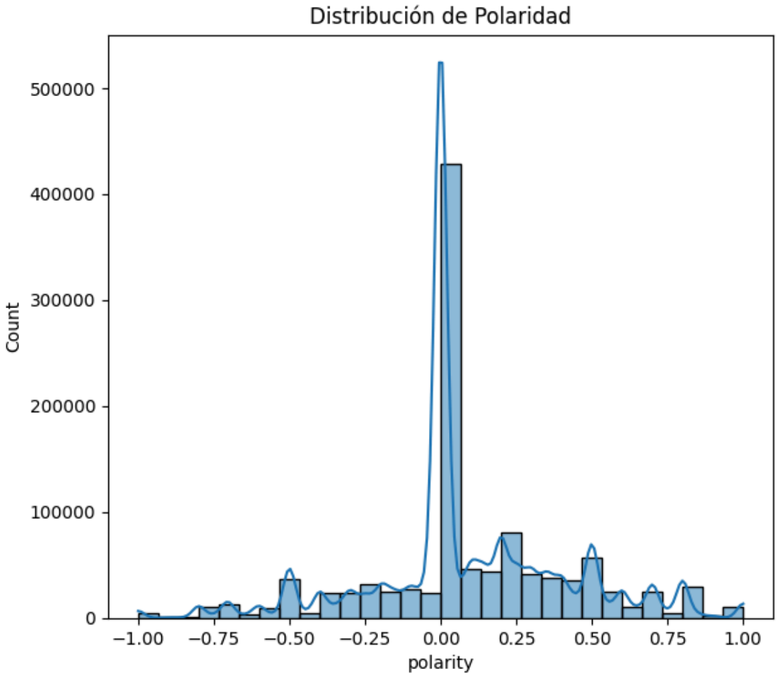


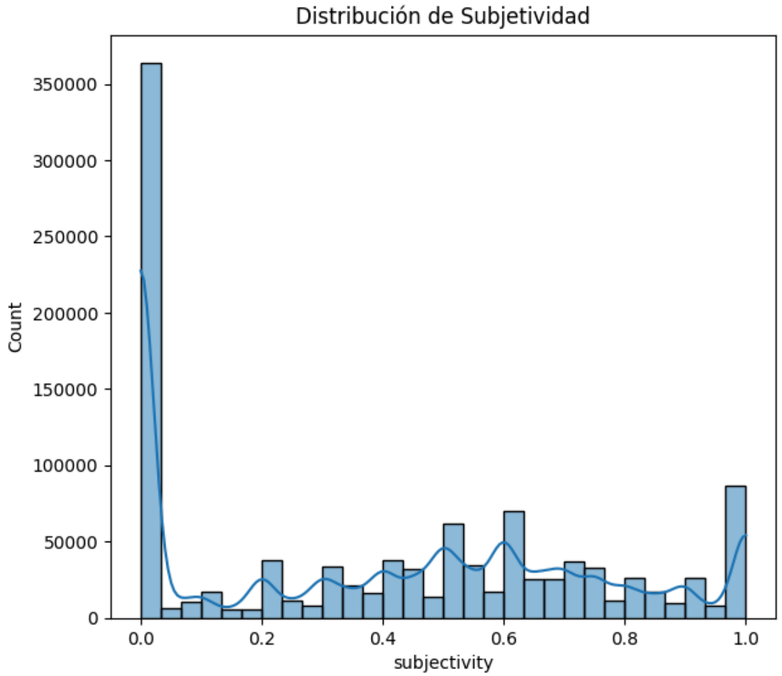
ANÁLISIS EXPLORATORIO DE DATOS (EDA)



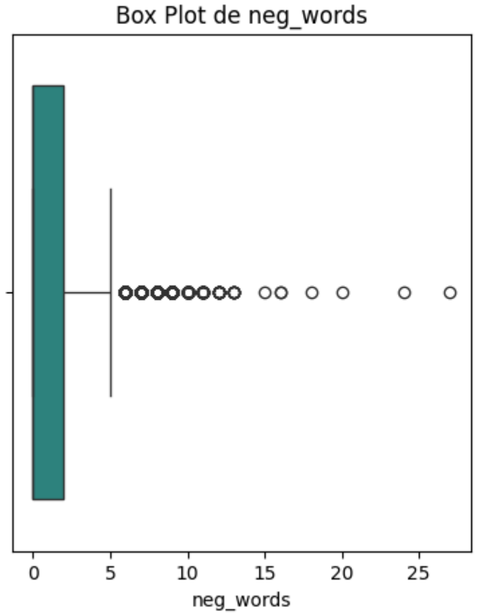


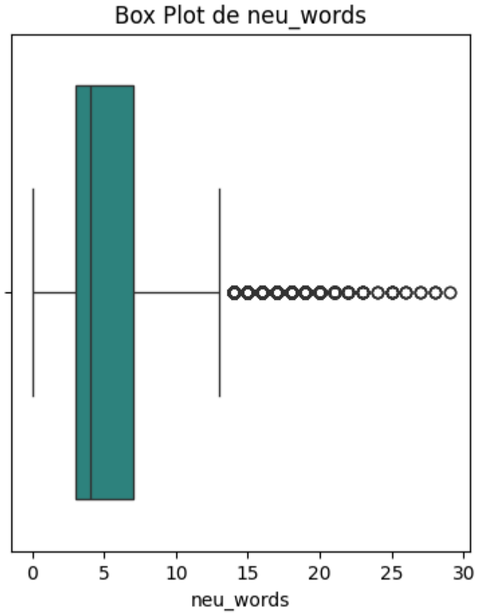


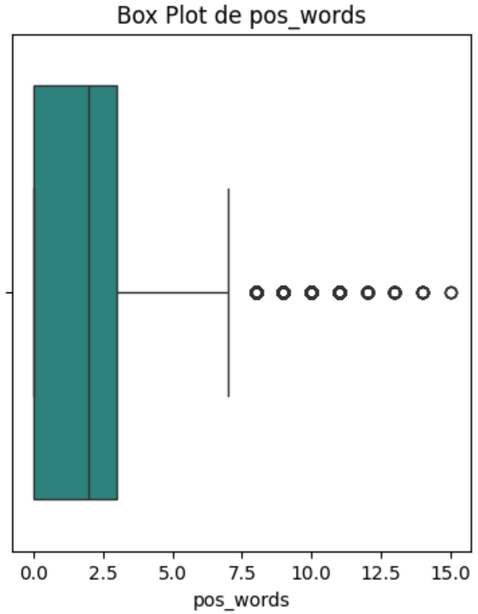


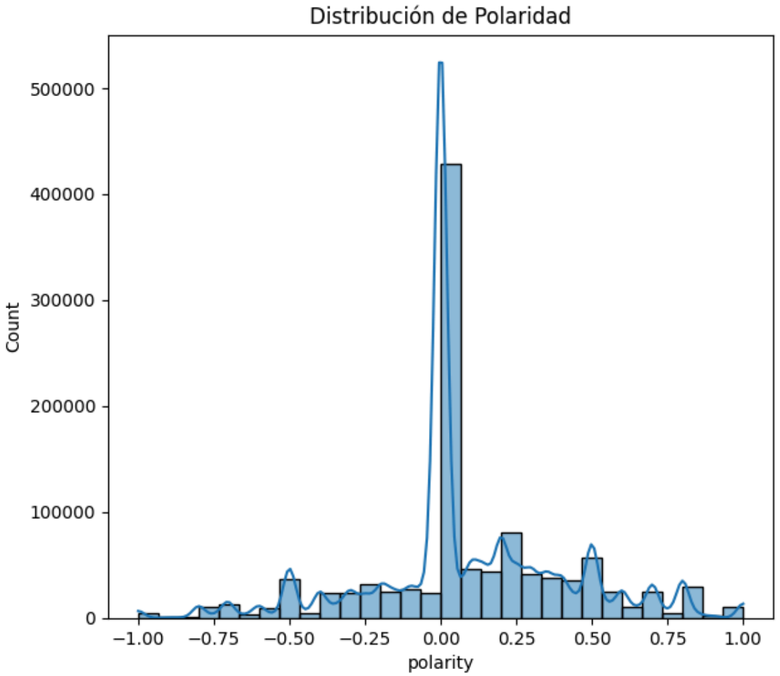


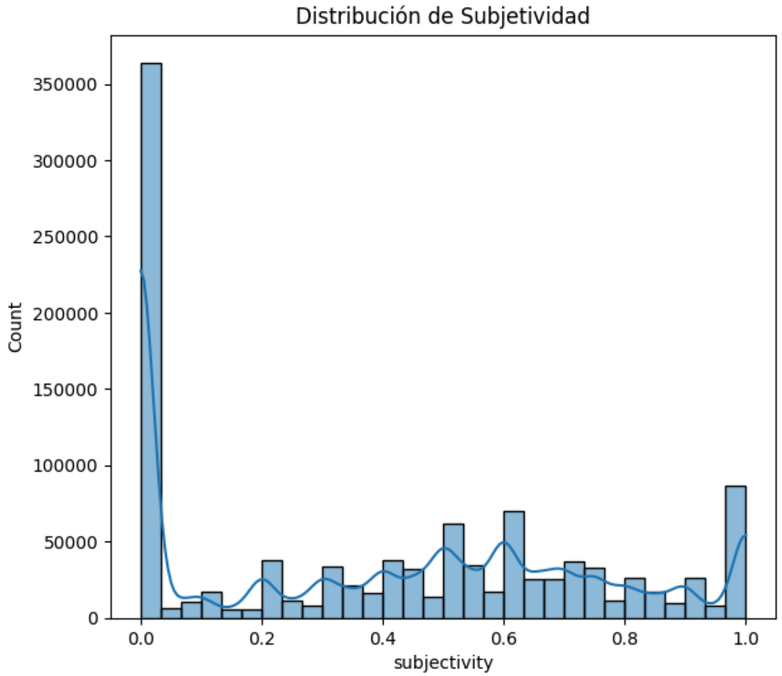
ANÁLISIS EXPLORATORIO DE DATOS (EDA)



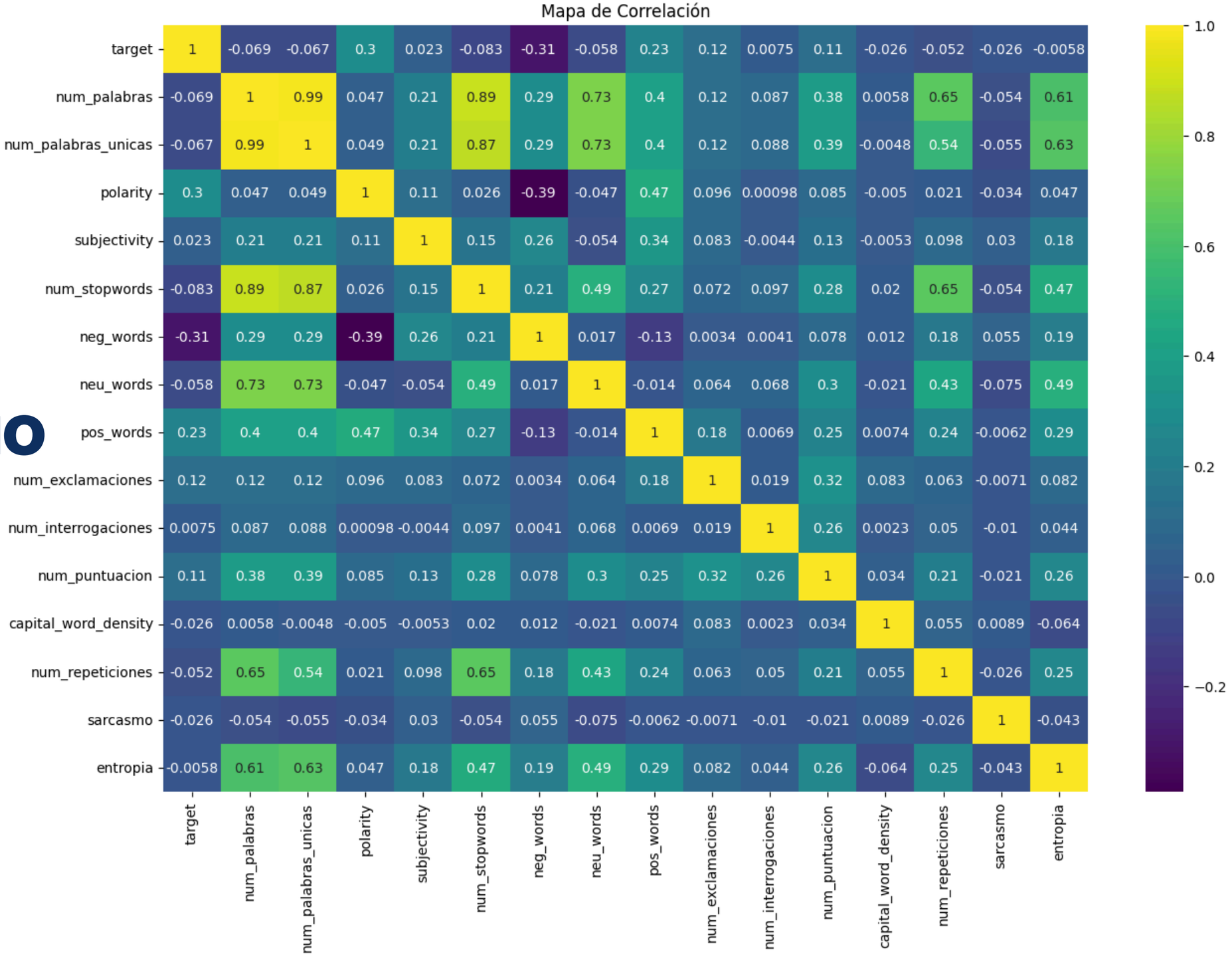




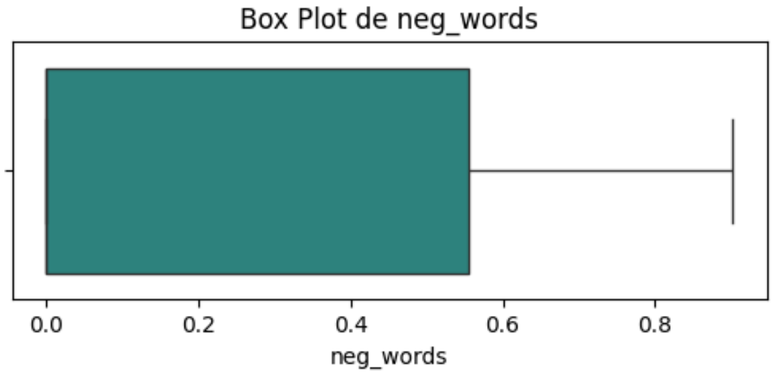


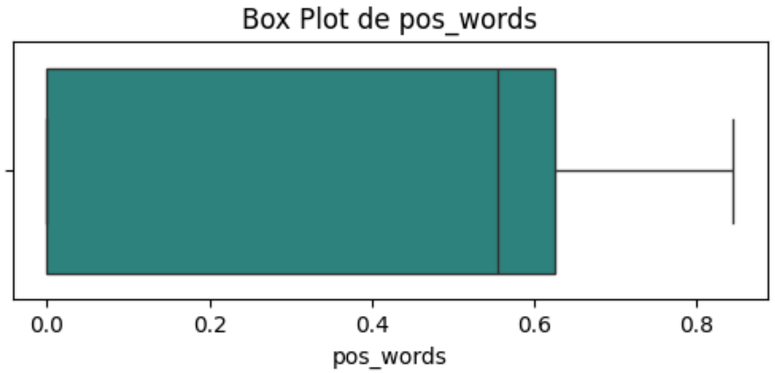


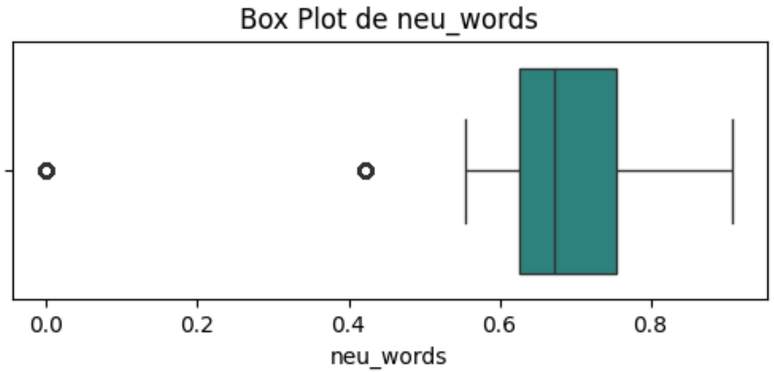
ANÁLISIS EXPLORATORIO DE DATOS (EDA)

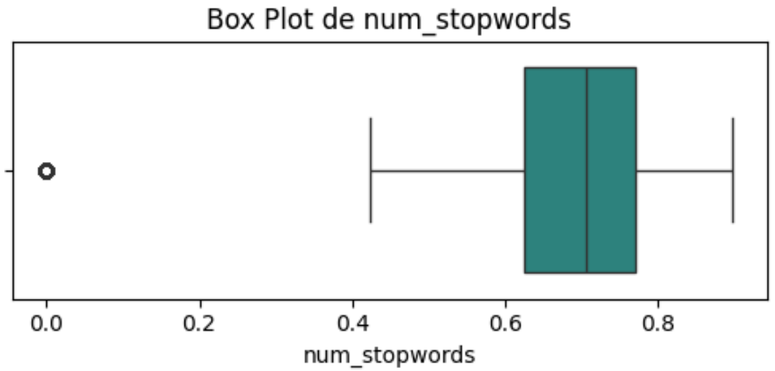


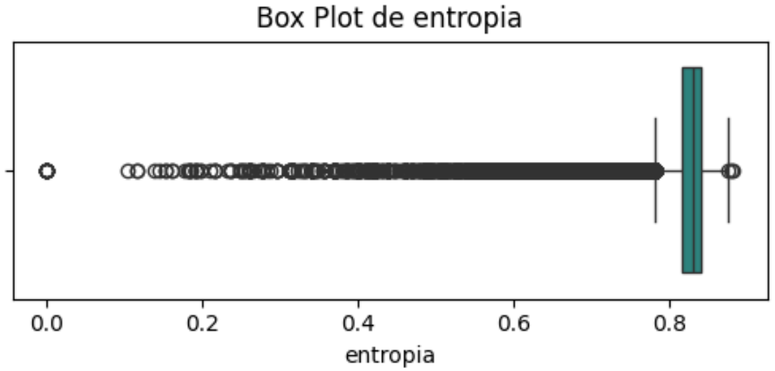
ANÁLISIS EXPLORATORIO DE DATOS (EDA)



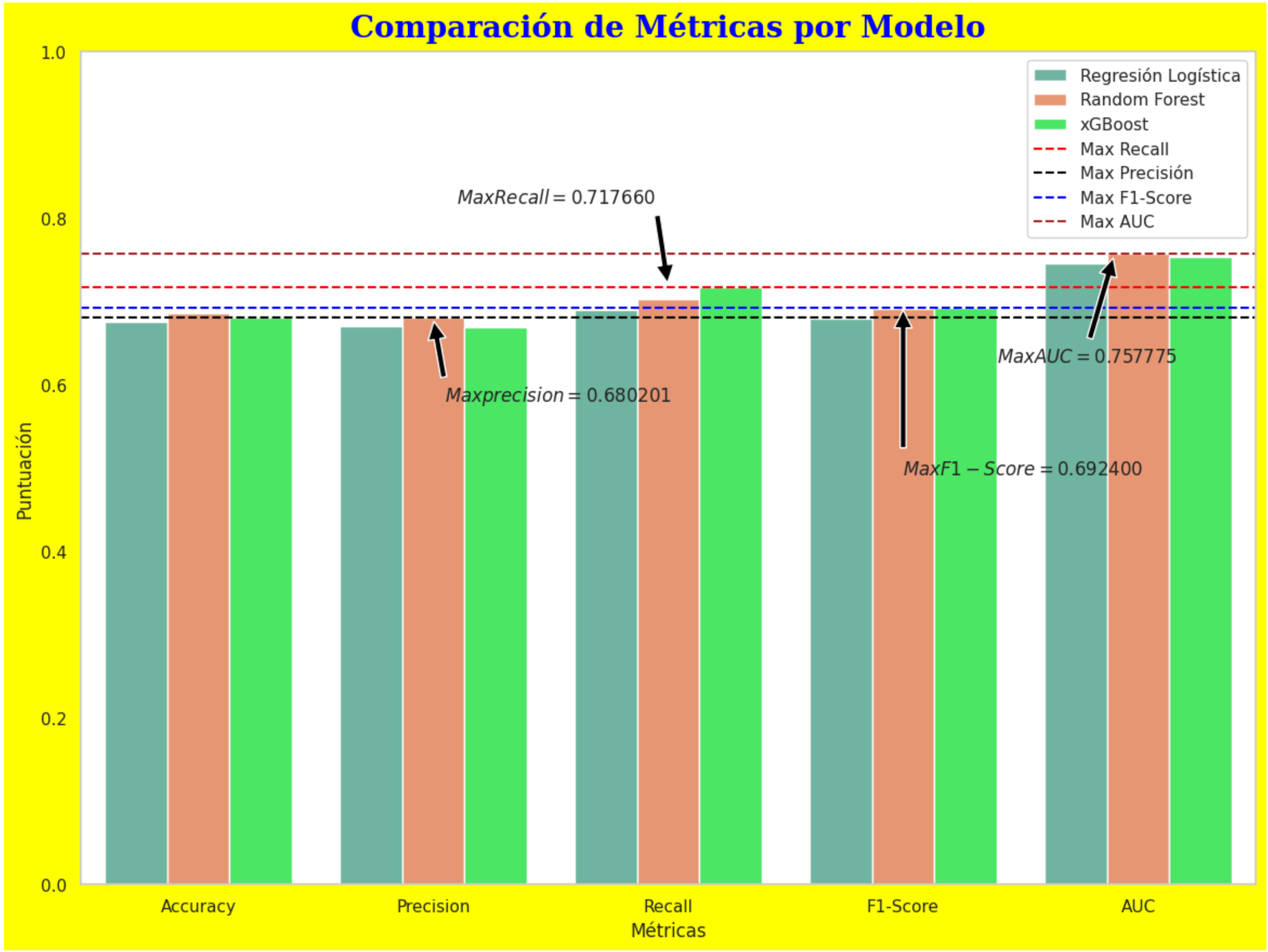




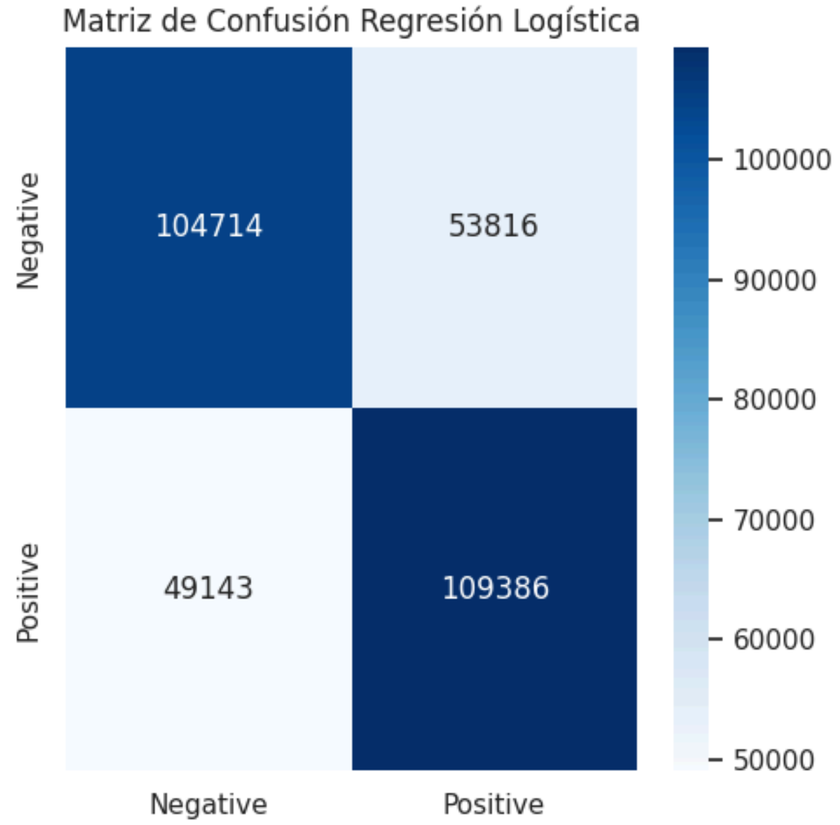


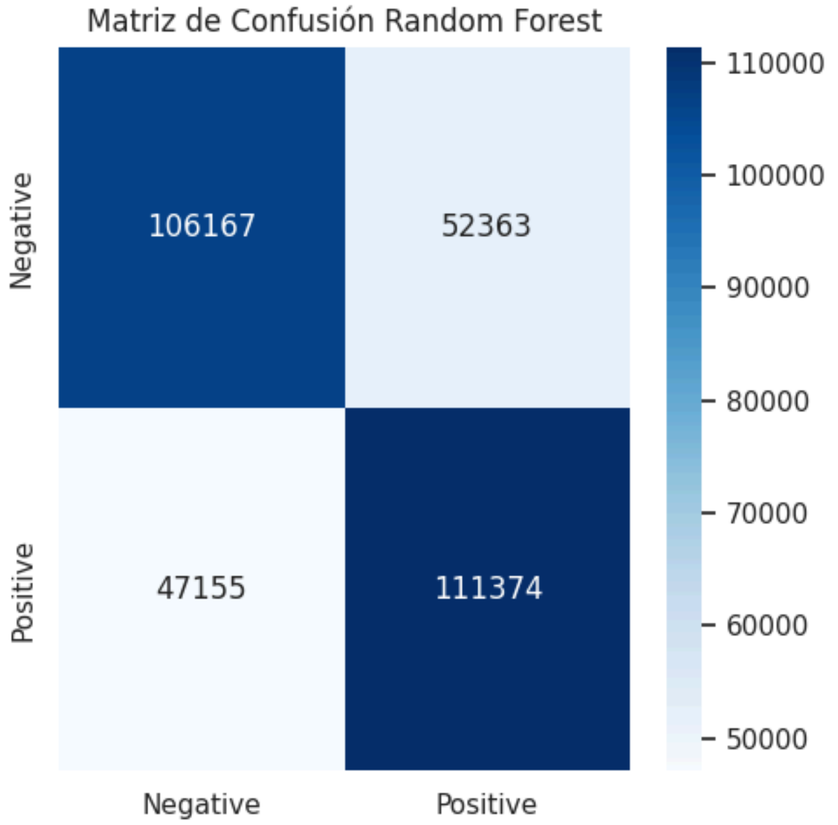


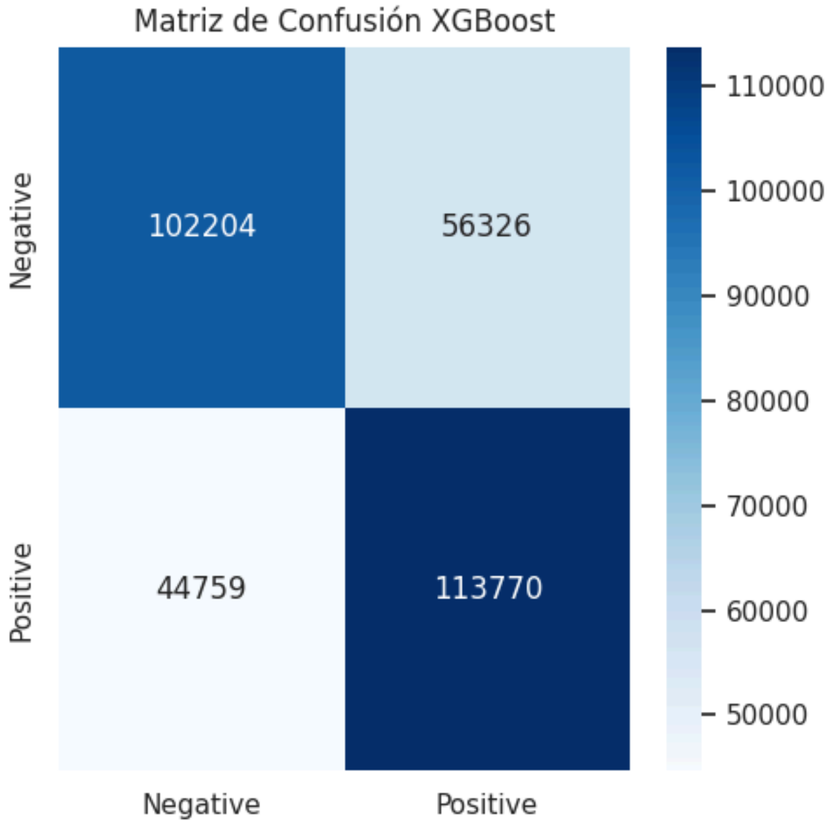
MODELADO DE LOS DATOS



MODELADO DE LOS DATOS







DESPLIEGUE DEL MODELO

APPLICATION WEB

El despliegue de la aplicación se realiza en Streamlit, utilizando un modelo de clasificación basado en XGBoost. A través de esta interfaz, los usuarios pueden ingresar tweets y obtener una predicción sobre su sentimiento, ya sea positivo o negativo.

CONCLUSIONES

- Se identificó que los cuatro usuarios con más tweets tienden a repetir mensajes, posiblemente indicando la presencia de bots. Para reducir sesgos, se limitó la cantidad de tweets por usuario.

- Se aplicó un tratamiento exhaustivo del dataset, eliminando datos nulos, tweets duplicados y ruido en el texto. Además, se eliminaron las stopwords para mejorar la calidad de los datos para la generación de características.

- Se crearon variables basadas en características del texto, como conteo de palabras, frecuencia de stopwords, signos de puntuación, sarcasmo, polaridad, subjetividad, entre otras.

- Se evaluaron 3 modelos de clasificación, Regresión Logística, Random Forest y XGBoost, analizando sus métricas claves como Recall, Precisión, F1-Score y AUC.

- Se seleccionó el modelo XGBoost, consideramos como la mejor opción para el despliegue, ya que obtuvo el mejor Recall (0.717660) y F1-Score (0.692400), si bien el AUC es ligeramente inferior al valor máximo (0.757775), consideramos que el modelo nos muestra un balance óptimo entre identificación correcta de la clase positiva y minimización de falsos negativos.

- Se implementó el modelo final en una aplicación Streamlit, permitiendo realizar predicciones de manera eficiente y visualmente accesible.