

UNIVERSIDAD DEL VALLE DE GAUATEMALA

Proyecto Final

Comparación de Clasificadores de SPAM/HAM

Javier Aguilar – 20611

Wilfredo Gallegos – 20399

Seminario Matemática /Alan Reyes-Figueroa

Recopilación de los datos

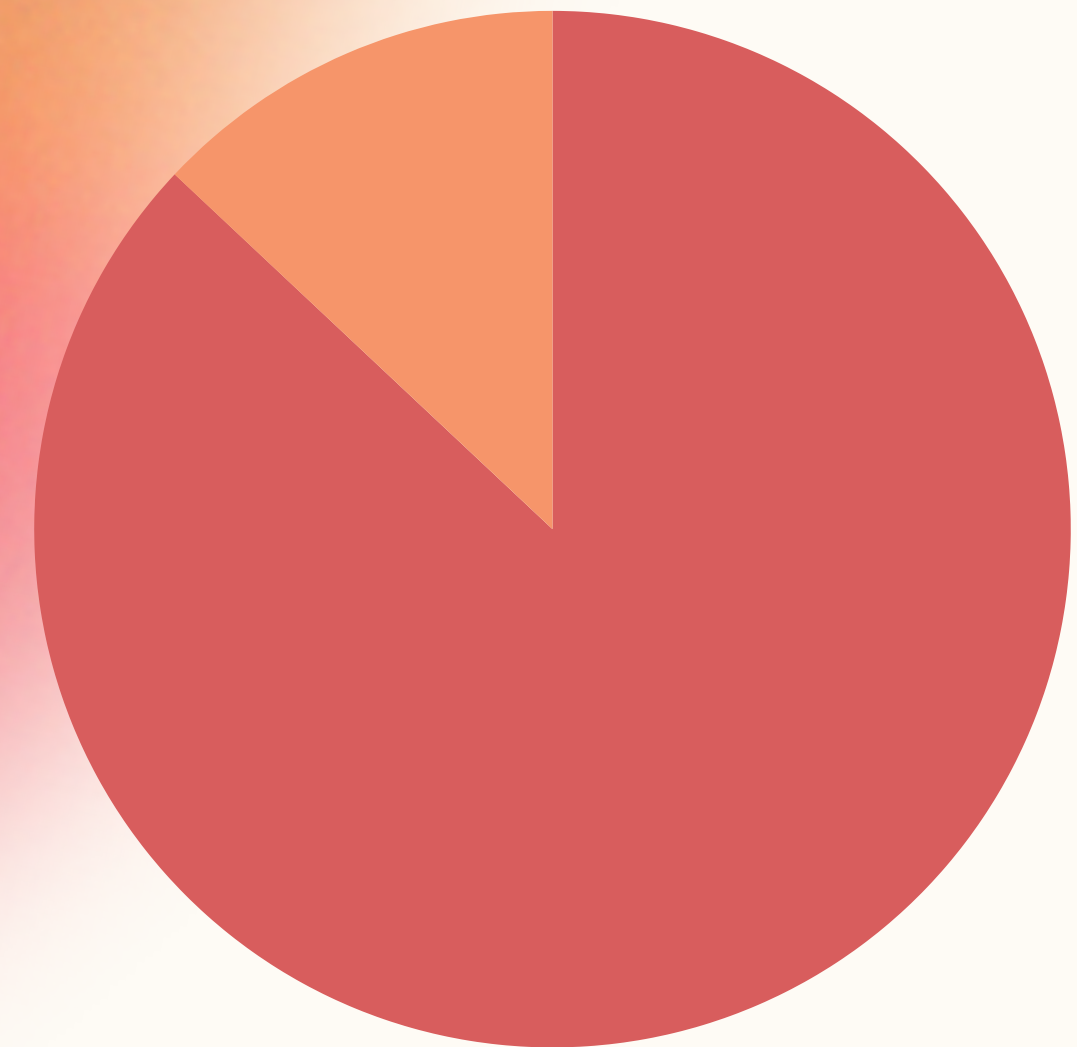


Kaggle

Se extrajo la base de datos de SPAM/HAM de la pagina oficial de Kaggle.com

- **Cantidad total de mensajes: 5157**

SPAM
13%



HAM
87%

Pre-procesamiento de los datos

- **Conversión a minúsculas:** Todas las letras se convierten a minúsculas para asegurar la consistencia en el texto.
- **Eliminación de caracteres no deseados:** Se eliminan todos los caracteres que no sean letras o espacios. Esto incluye números, símbolos de puntuación, etc.
- **Eliminación de palabras vacías:** Se eliminan las palabras comunes que no aportan mucho significado al análisis de texto.
- **Aplicación a los datos:** Finalmente, estos pasos de limpieza se aplican a cada mensaje en la base de datos.

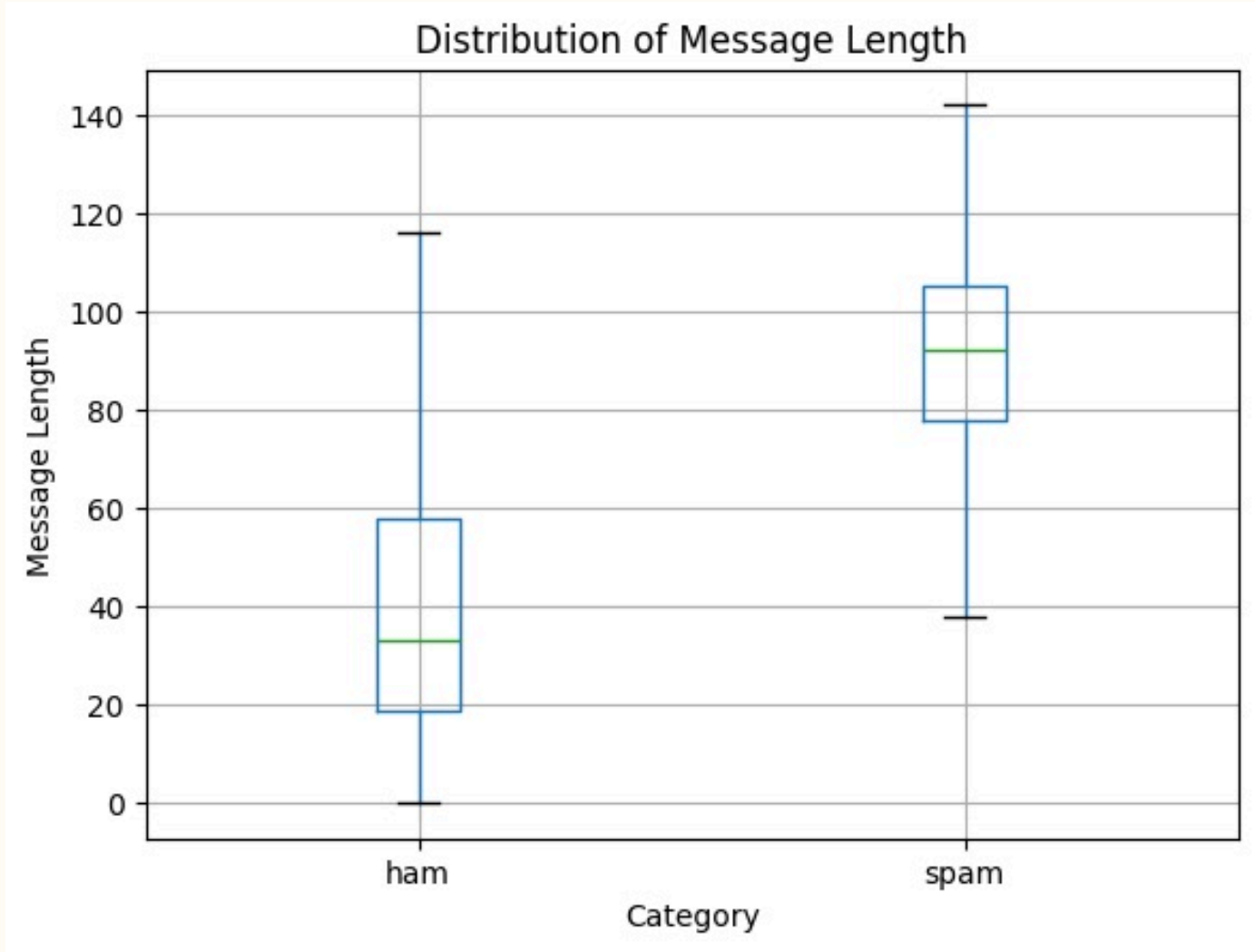
Mensaje original

hi its me you are probably having
too much fun to get this message
but i thought id txt u cos im bored!
and james has been farting at me all
night

Mensaje despues de la limpieza

hi probably much fun get message
thought id txt u cos im bored james
farting night

Análisis Exploratorio



Category	Palabras únicas
HAM	3483
SPAM	879

Palabras en común ambas categorías

964

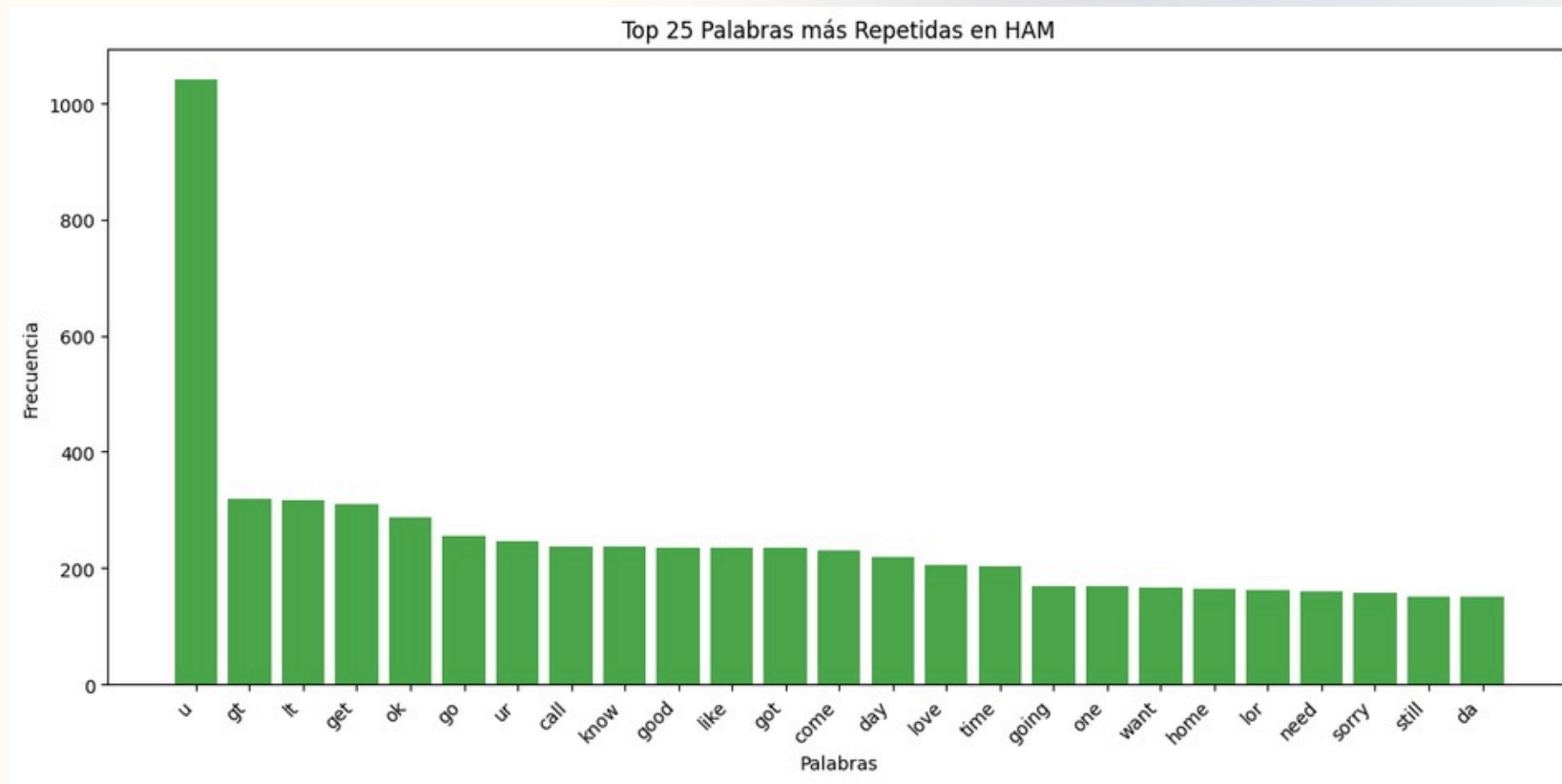
Nubes

[illegible]

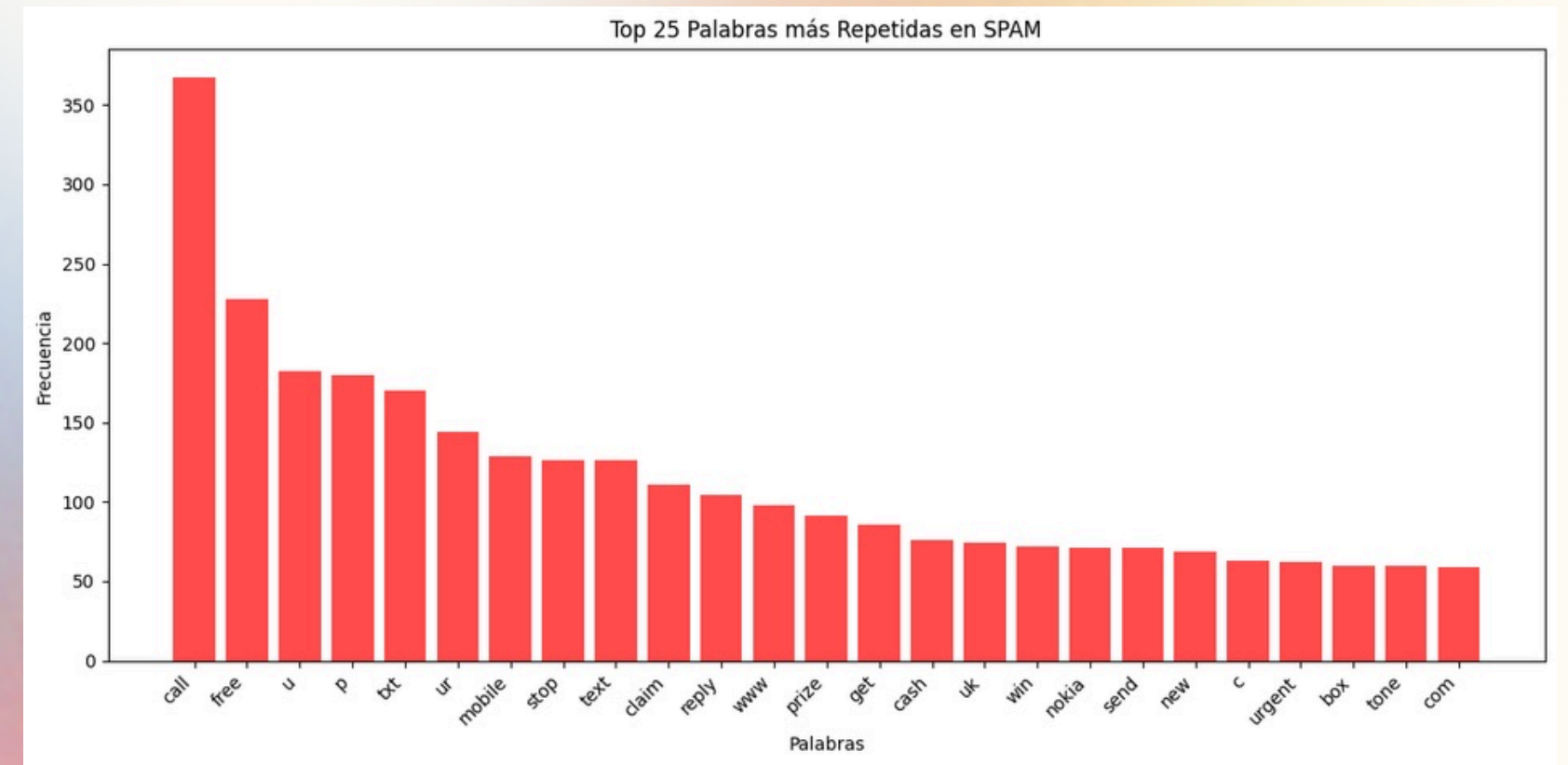
Palabras

[illegible]

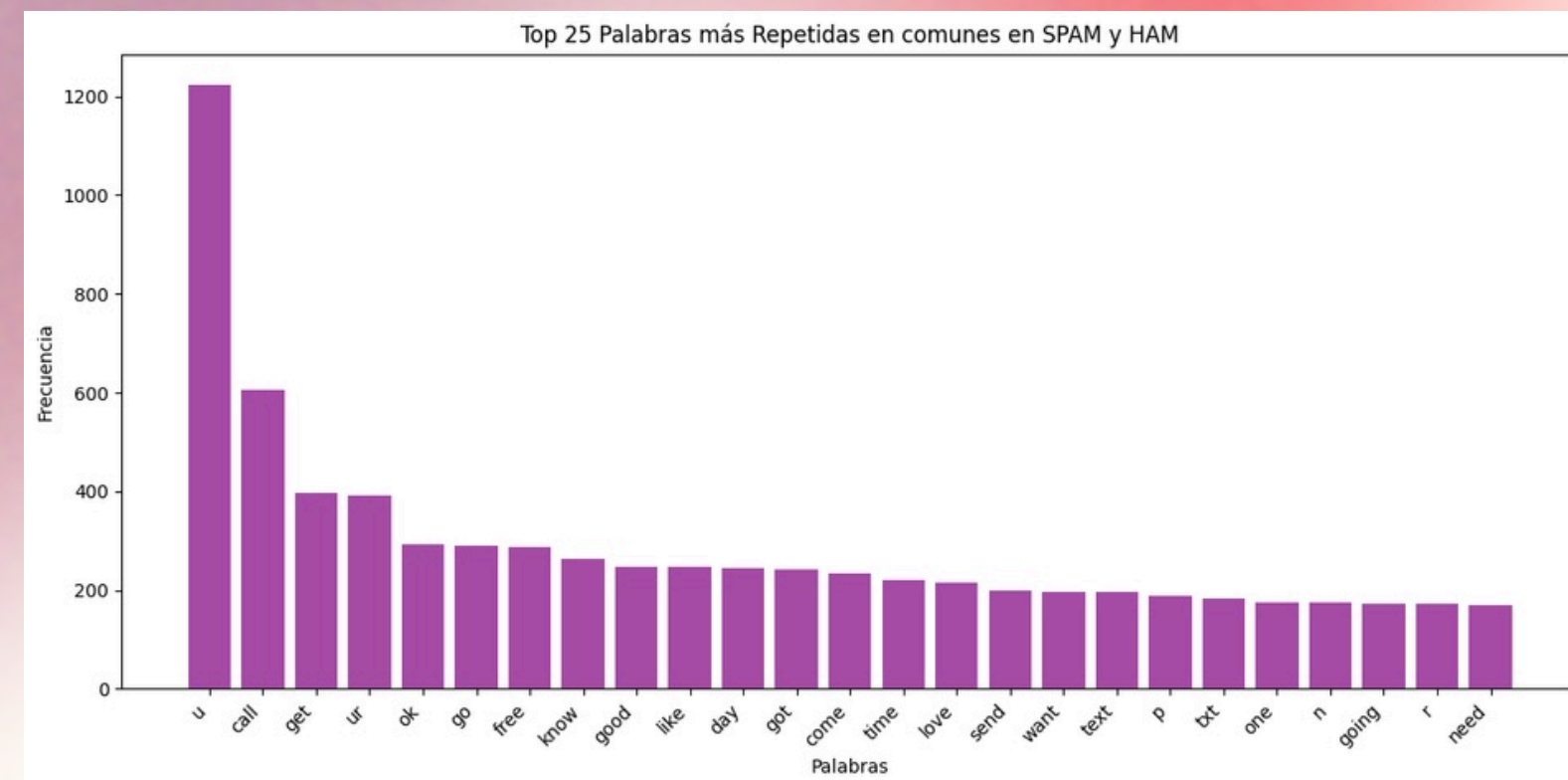
Top 25 palabras más repetidas en HAM



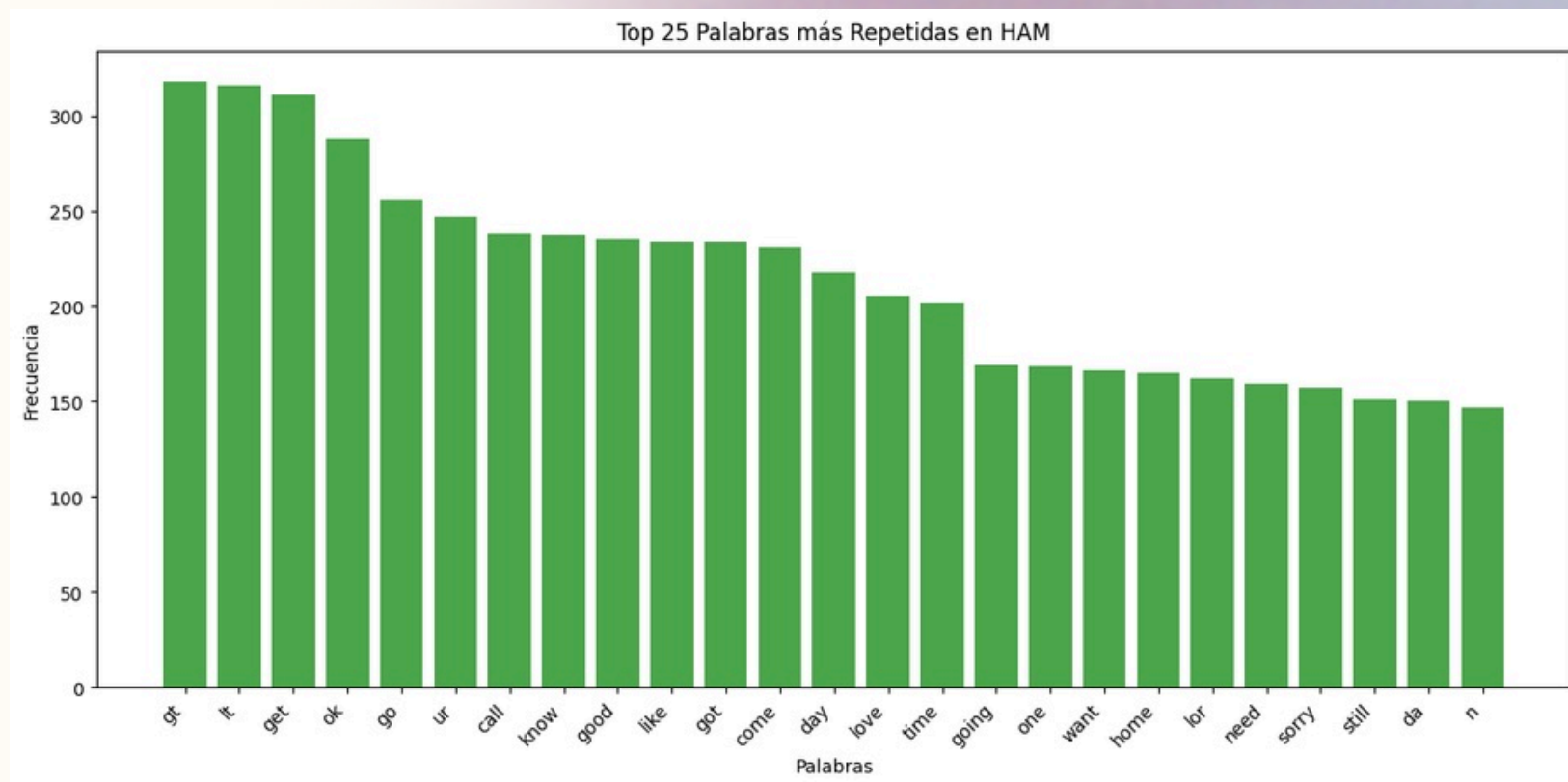
Top 25 palabras más repetidas en SPAM



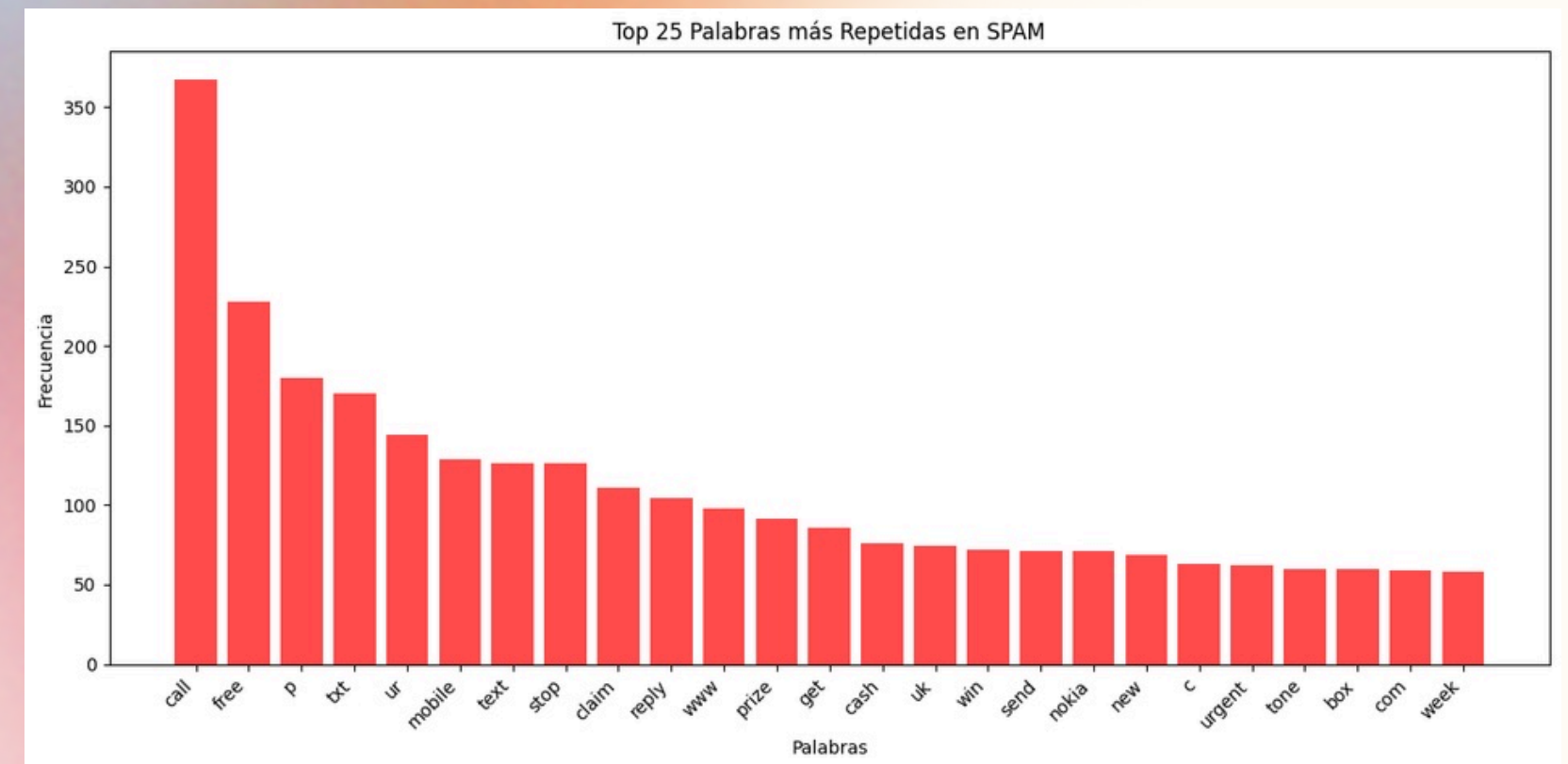
Top 25 palabras más repetidas en común



Top 25 palabras más repetidas en HAM



Top 25 palabras más repetidas en SPAM



	Category	Message	v2_clean	v2_len
0	ham	Go until jurong point, crazy.. Available only ...	go jurong point crazy available bugis n great ...	82
1	ham	Ok lar... Joking wif u oni...	ok lar joking wif oni	23
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	free entry wkly comp win fa cup final tkts st ...	101
3	ham	U dun say so early hor... U c already then say...	dun say early hor c already say	35
4	ham	Nah I don't think he goes to usf, he lives aro...	nah think goes usf lives around though	38
...	...	...	...	...
4452	ham	And that is the problem. You walk around in "j...	problem walk around julianaland oblivious goin...	158

Modelo Bayesiano

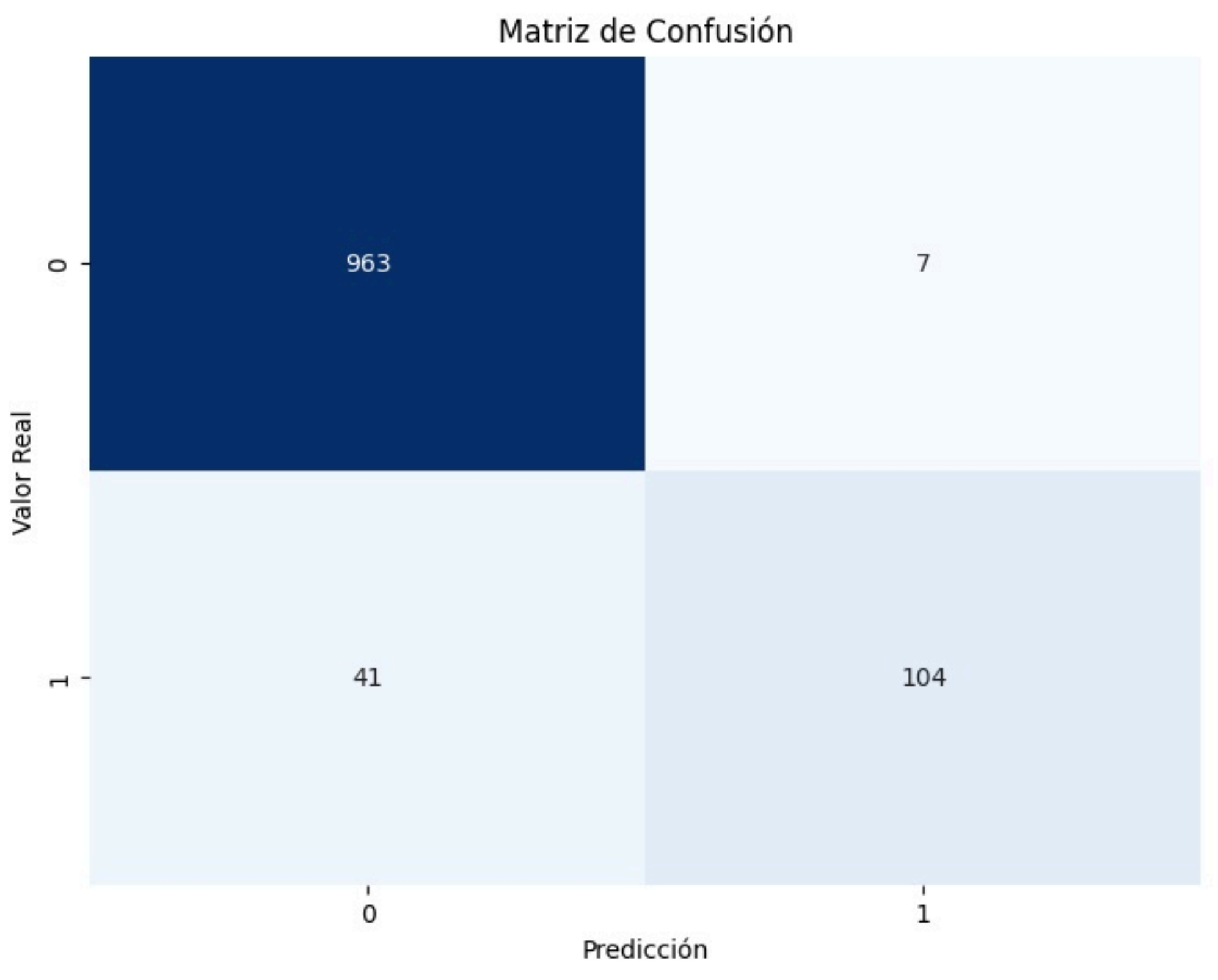
$$P(S|W) = \frac{P_1 P_2 \dots P_n}{P_1 P_2 \dots P_n + (1 - P_1)(1 - P_2) \dots (1 - P_n)}$$

Donde

$$P(W) = \frac{\frac{N_{w,s}}{N_s}}{\frac{N_{w,h}}{N_h} + \frac{N_{w,s}}{N_s}}$$

- $N_{w,s}$: Cantidad de correos Spam que contienen la palabra W
- $N_{w,h}$: Cantidad de correos Ham que contienen la palabra W
- N_s : Cantidad total de Correos Spam
- N_h : Cantidad total de correos Ham

Resultados modelo Bayesiano



Métrica	Valor (%)
Precision Score	93.69
Recall Score	71.72
F1 Score	81.25

Tiempo de ejecución: 0.1s

Resultados modelo LSTM

Resultados modelo LSTM

- **Características de LSTM:**

- Procesa secuencias de datos.
- Recuerda información a largo plazo.

- **Preparación de Datos:**

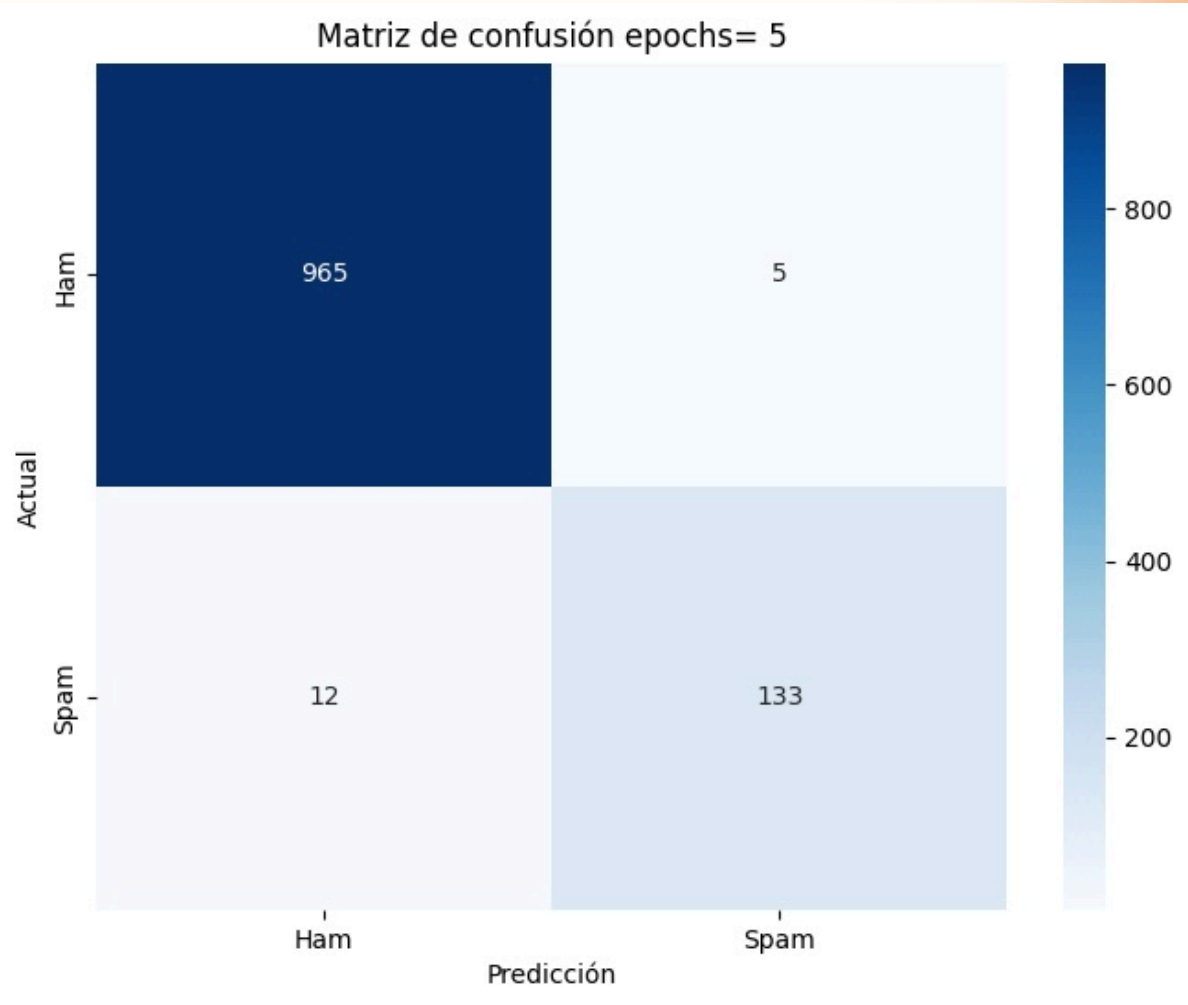
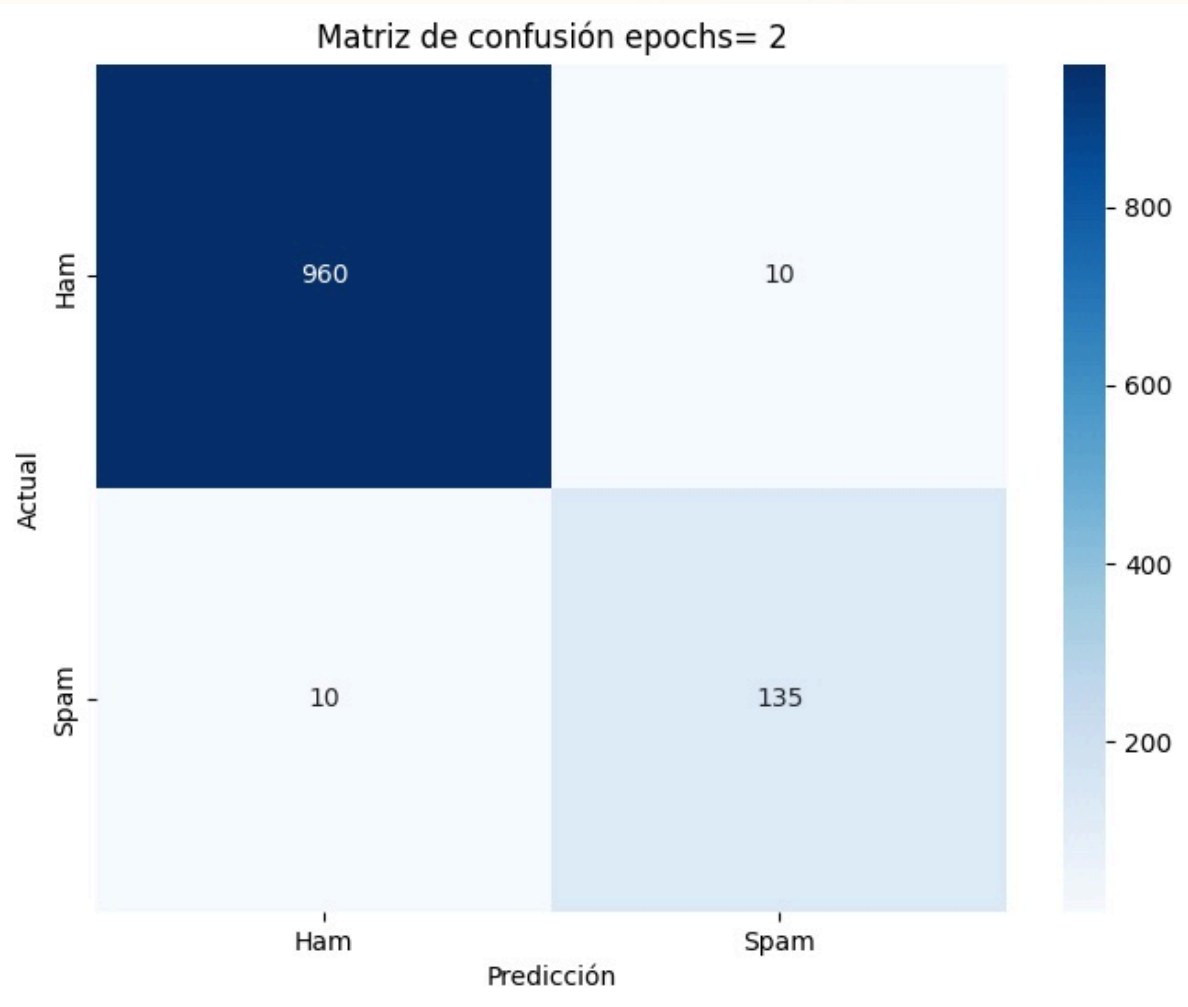
- Etiquetado de mensajes como spam o ham.
- Conversión de etiquetas en números.
- Tokenización de textos en palabras.
- Conversión de textos en secuencias numéricas.
- Aplicación de padding para uniformidad de secuencias.

- **Estructura del Modelo:**

- Una Capa que convierte secuencias en vectores de características.
- Dos capas de LSTM con 128 unidades cada una.
- Capas de dropout para evitar sobreajuste.

- **Entrenamiento y Evaluación:**

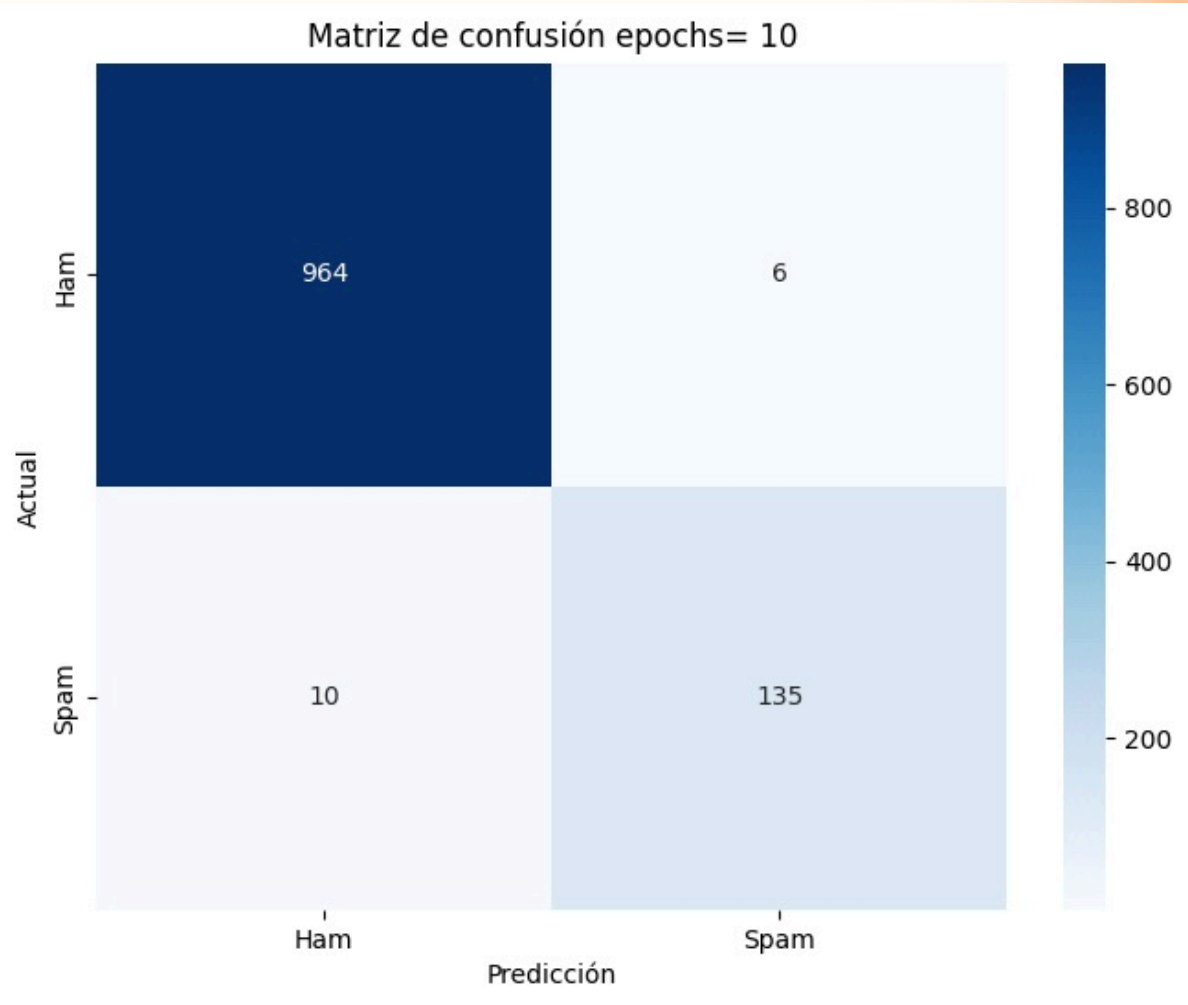
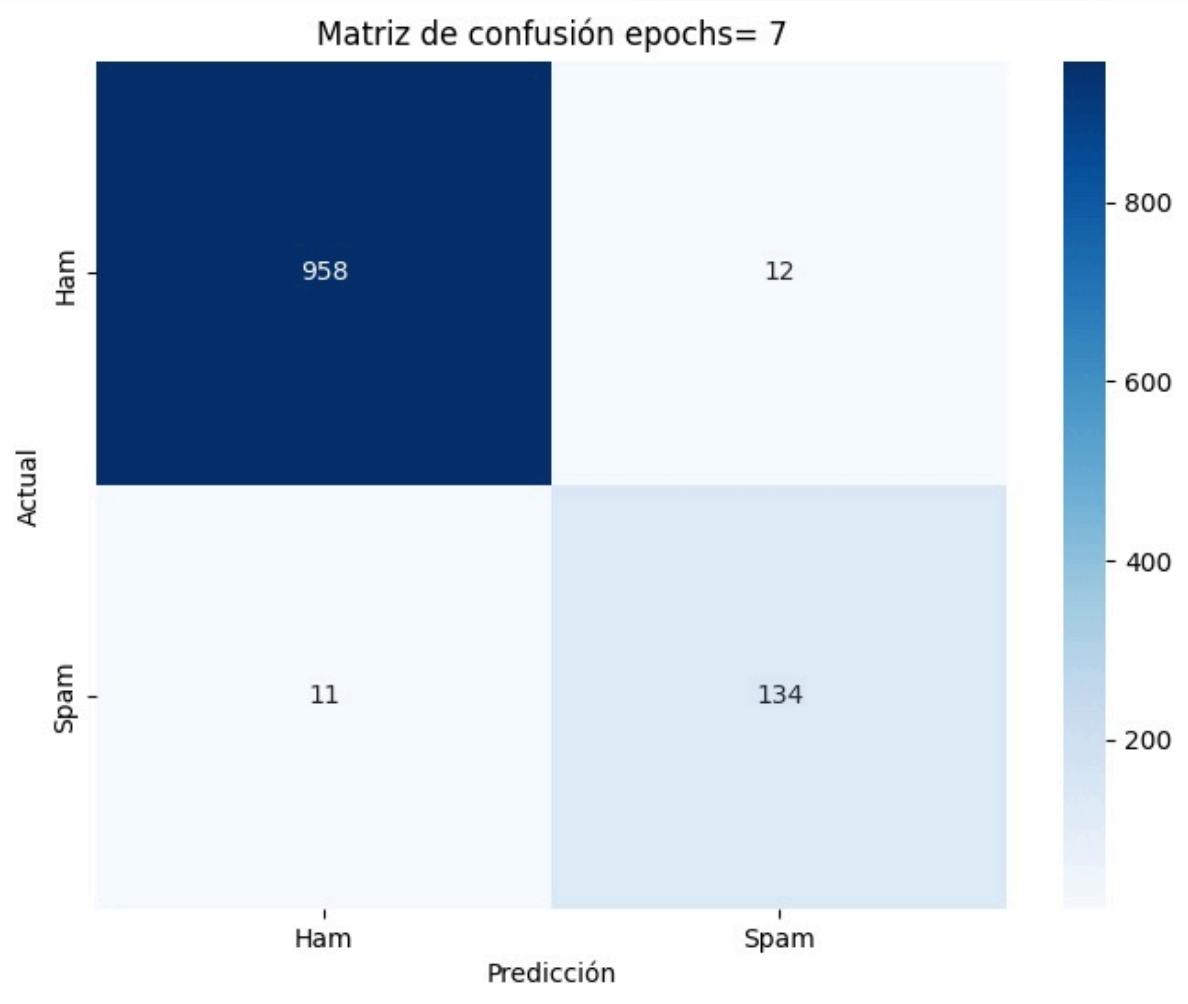
- Optimizador: Adam.
- Entrenamiento durante 10,7,5 y 2 épocas.
- Alta precisión en clasificación de mensajes en conjunto de prueba



Métrica	Valor (%)
Precision Score	93.10
Recall Score	93.10
F1 Score	93.10

Métrica	Valor (%)
Precision Score	96.38
Recall Score	91.72
F1 Score	93.99

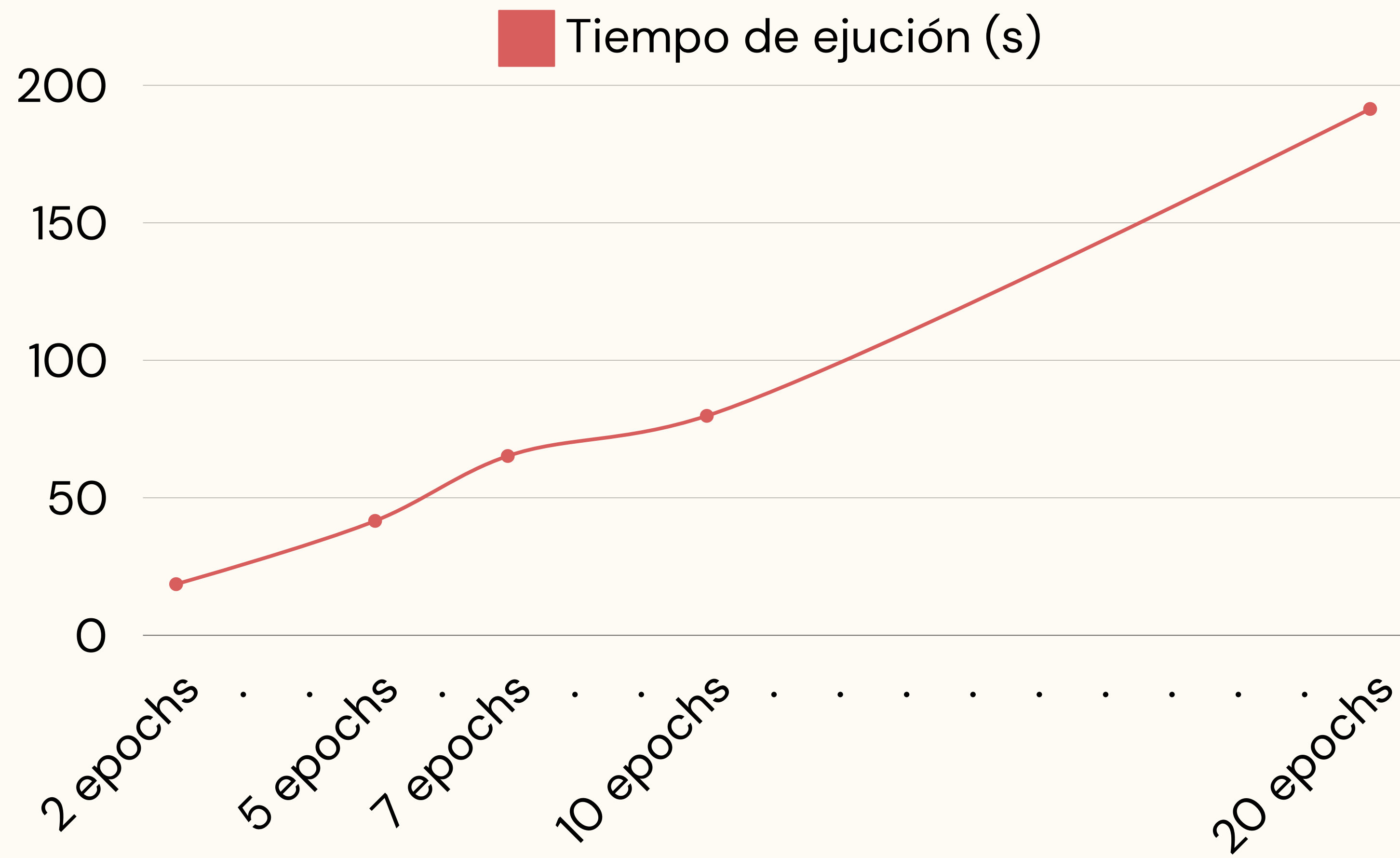
Matrices de
confusión para
Epochs 2 y 5



Métrica	Valor (%)
Precision Score	91.18
Recall Score	92.41
F1 Score	91.79

Métrica	Valor (%)
Precision Score	95.74
Recall Score	93.10
F1 Score	94.40

Matrices de confusión para Epochs 7 y 10



Conclusiones y recomendaciones

Si se quiere optimización de recursos sacrificando ligeramente los resultados, se recomienda el uso de modelo Bayesiano.

Si se busca tener un resultado mas preciso, pero utilizando mas recursos se recomienda el uso del modelo LSTM

Para la predicción de spam/ham, no es necesario realizar tantas iteraciones para el modelo lstm, ya que se obtienen buenos resultados con un bajo entrenamiento.

Se recomienda entrenar los modelos con un conjunto de datos mas diverso, para poder hacer predicción sobre correos mas variados.

¡Gracias por su atención.

¿Preguntas?