Universidad del Valle de Guatemala
Inteligencia Artificial

Proyecto 02

Carlos Edgardo López Barrera 21666

Guatemala, 06 de mayo del 2024

Análisis Exploratorio

El dataset que describes parece ser una colección de mensajes de texto clasificados como "ham" y "spam".

Estructura del Dataset:

Columna v1: Indica la etiqueta de cada mensaje, con valores "ham" para mensajes legítimos y "spam" para mensajes no deseados.

Columna v2: Contiene el texto del mensaje. Los textos varían desde conversaciones cotidianas y solicitudes hasta anuncios promocionales y estafa

Descripción del Contenido:

Mensajes Ham: Estos mensajes incluyen interacciones normales como conversaciones amistosas, coordinaciones diarias, expresiones de emociones, y otras comunicaciones personales o triviales. Ejemplos incluyen mensajes sobre planes, emociones, y respuestas cotidianas.

Mensajes Spam: Estos textos suelen incluir ofertas comerciales, concursos, promociones, o alertas que implican algún tipo de ganancia o incentivo que generalmente es falso. Comúnmente, estos mensajes contienen instrucciones para responder o llamar a números de teléfono, a menudo asociados con costos adicionales o suscripciones automáticas.

Gráficas y/o reportes

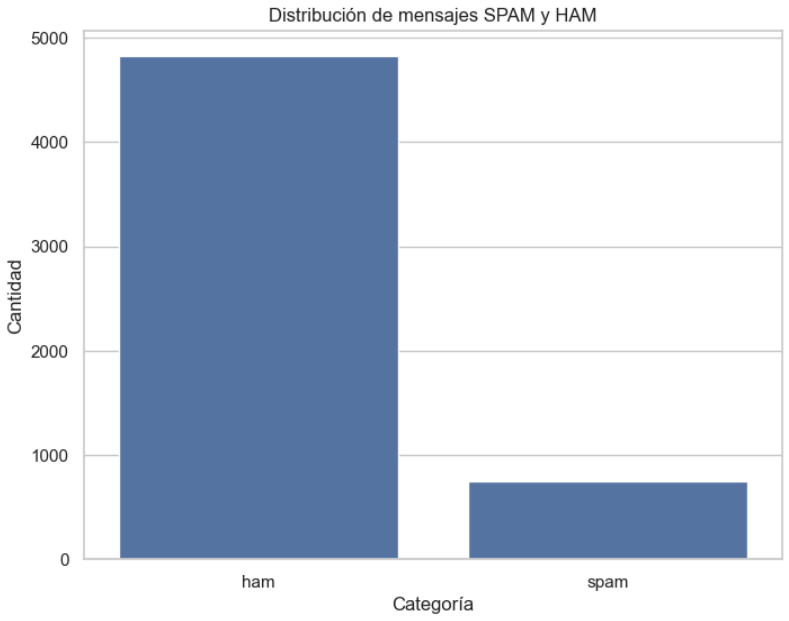

Figura #1 - Distribución de mensajes SPAM y HAM

La Figura #1 muestra la distribución de mensajes clasificados como "ham" y "spam" en un conjunto de datos. Los principales hallazgos o información que se pueden deducir de esta gráfica son: La cantidad de mensajes "ham" es significativamente mayor que la cantidad de mensajes "spam". Esto indica que en el conjunto de datos analizado, los mensajes legítimos predominan sobre los mensajes no deseados.

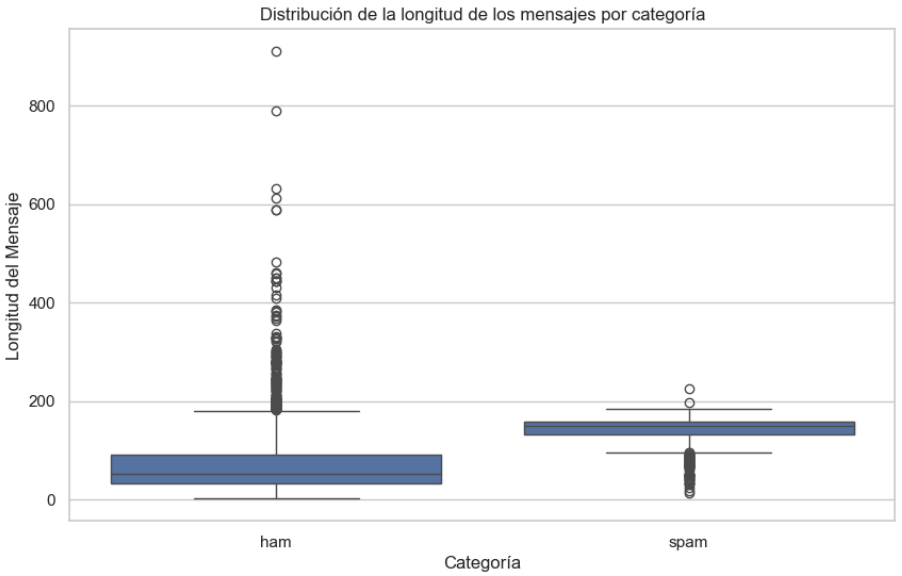

Figura #2 - Distribución de la longitud de los mensajes por categoría

La Figura #2 muestra un boxplot que compara la distribución de la longitud de los mensajes categorizados como "ham" y "spam". A partir de esta visualización, podemos deducir varios puntos clave sobre las características de los mensajes en cada categoría: Los mensajes "ham" tienden a ser más cortos en comparación con los mensajes "spam". Esto se puede observar en la longitud de las cajas, donde la caja del "ham" es más corta y está más cercana al eje de origen. La mediana para los mensajes "ham" es considerablemente más baja que la de los mensajes "spam", lo que indica que la longitud típica de un mensaje “ham” es menor.
La variabilidad en la longitud de los mensajes "ham" es menor comparada con la de los mensajes "spam". Esto se refleja en las cajas más estrechas para los mensajes "ham".
Los mensajes "spam" no solo tienen una mediana más alta, sino que también muestran una mayor dispersión en longitud, indicada por la mayor altura de su caja en el gráfico.
Los mensajes "spam" tienen menos valores atípicos en comparación con los "ham", pero estos valores atípicos tienden a alcanzar longitudes más extremas, lo que sugiere que hay algunos mensajes "spam" mucho más largos que el promedio.

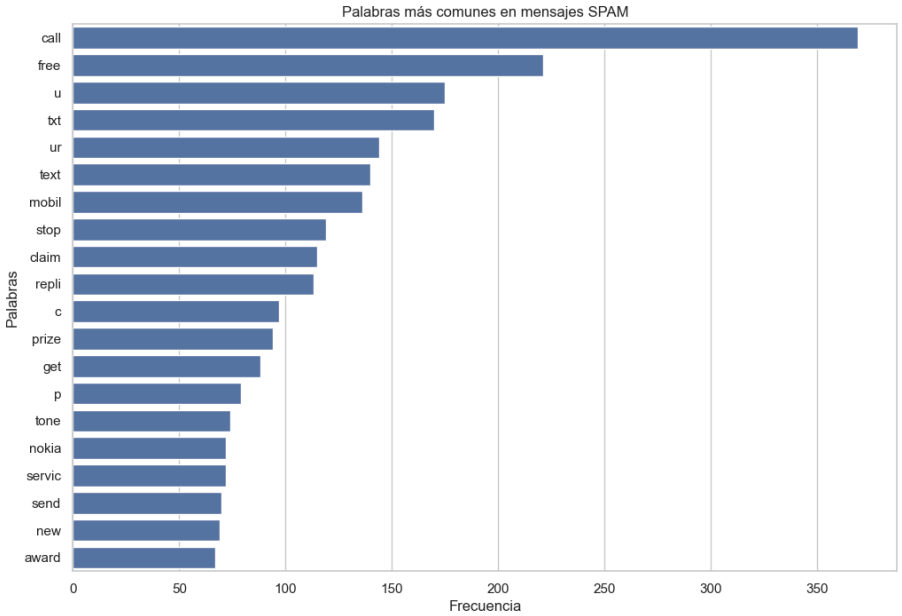

Figura #3 - Palabras más comunes en SPAM

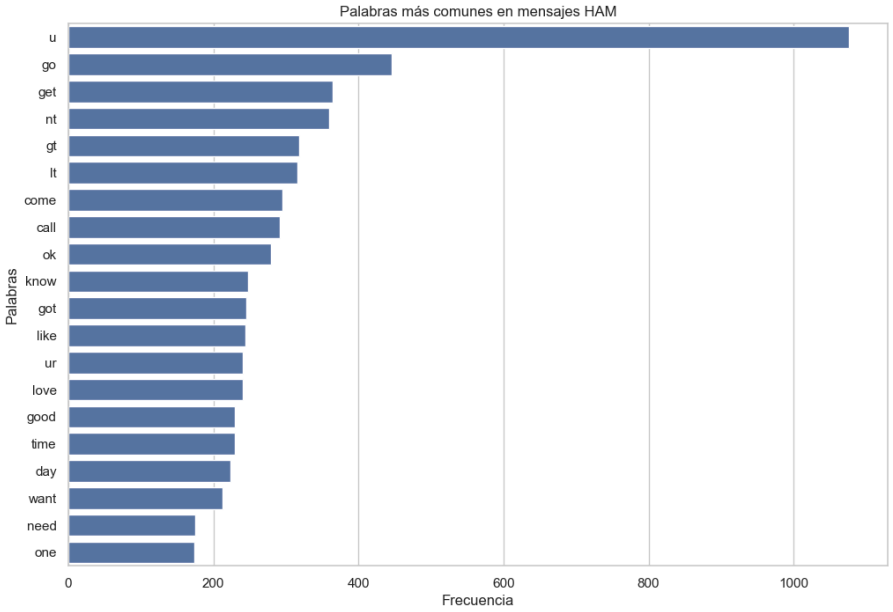

Figura #4 - Palabras más comunes en HAM

Las figuras #3 y #4, muestran las palabras más comunes en mensajes clasificados como "spam" y "ham", respectivamente, junto con sus frecuencias de aparición.
Palabras como "call", "free", "u", "txt", y "ur" son especialmente prevalentes en los mensajes de spam. Esto sugiere que estos mensajes a menudo contienen llamados a la acción o ofertas que se comunican de manera directa y que utilizan un lenguaje imperativo o promocional.
Ahora bien, palabras como "u", "go", "get", y "nt" son comunes, pero su contexto es probablemente diferente al de los mensajes de spam.

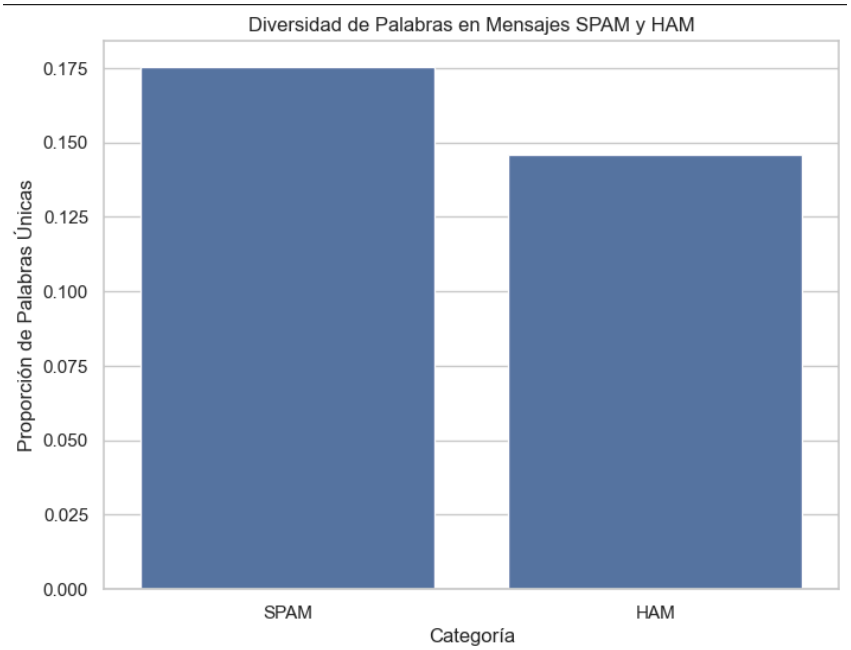

Figura #5 - Diversidad de palabras en mensajes SPAM y HAM

La figura #5 compara la diversidad de palabras en mensajes clasificados como "SPAM" y "HAM". Esto refleja cuán variado es el vocabulario utilizado en cada tipo de mensaje.
Los mensajes de spam muestran una proporción ligeramente más alta de palabras únicas comparada con los mensajes ham. Esto podría indicar que los mensajes de spam utilizan un rango más amplio de palabras para atraer la atención de los destinatarios.

Limpieza de datos

Se han utilizado varias funciones esenciales para transformar el texto bruto en una forma más estandarizada y útil para el análisis posterior.

- Verificación de Tipo
  - Asegura que la entrada sea una cadena de texto. Esto previene errores durante la manipulación de texto que podría no ser un string.
- Conversión a Minúsculas
  - Convertir todo el texto a minúsculas para unificar el caso de las letras, eliminando diferencias entre mayúsculas y minúsculas.
- Tokenización
  - Dividir el texto en palabras o "tokens", lo que facilita el procesamiento individual de cada término.
- Eliminación de Puntuación y Números
  - Remover caracteres no alfabéticos, incluyendo puntuación y números, de cada token.
- Eliminación de Palabras Vacías

    - Descartar palabras que generalmente no contribuyen al significado del texto (stopwords) y tokens que no sean completamente alfabéticos.
- Lematización
    - Convertir las palabras a su forma de base o lema, tratando diferentes formas de una palabra como una sola (por ejemplo, "running" a "run").
- Stemming
    - Reducir las palabras a su raíz o "stem", que no necesariamente tiene que ser una palabra válida.

Modelo

- Paso 1: Cálculo de Frecuencias de Palabras
  La función calculate_word_frequencies es utilizada para calcular la frecuencia de cada palabra en los mensajes clasificados como SPAM y HAM.

  spam_messages y ham_messages: Segmentación del DataFrame en dos grupos según la etiqueta ('spam' o 'ham').

  spam_words y ham_words: Extracción de todas las palabras de los mensajes de cada grupo.

  spam_word_count y ham_word_count: Conteo de la frecuencia de cada palabra en los mensajes de spam y ham, respectivamente.

  total_spam y total_ham: Conteo del número total de mensajes en cada categoría.

- Paso 2: Cálculo de Probabilidades
  La función calculate_probabilities utiliza los conteos obtenidos para calcular la probabilidad de que un mensaje sea SPAM dado que contiene ciertas palabras (usando Bayes)

- Paso 3: Predicción de SPAM o HAM
  La función predict_spam se utiliza para predecir si un mensaje dado es SPAM o HAM basado en las probabilidades calculadas para sus palabras.

  words: Lista de palabras obtenidas de limpiar el texto del mensaje.

  recognized_words: Palabras del mensaje que están en el diccionario de probabilidades.

  probs: Lista de probabilidades de SPAM para cada palabra reconocida.

spam_prob: Calculada como la probabilidad de que el mensaje completo sea SPAM basándose en la combinación de las probabilidades de sus palabras.

Es importante mencionar que para el cálculo de la probabilidades se utilizaron las siguientes fórmulas:

- Probabilidad de que un texto sea SPAM dado que contiene la palabra W:

$$P(S|W) = \frac{P(W|S)P(S)}{P(W|S)P + P(W|H)P(H)}$$

P(S|W) es la probabilidad de que un texto sea SPAM dado que contiene la palabra W.
P(W|S) es la probabilidad de la palabra W aparezca en un texto que es SPAM.
P(W|H) es la probabilidad de la palabra W aparezca en un texto que es HAM.
P(S) es la probabilidad de que cualquier texto sea SPAM.
P(H) es la probabilidad de que cualquier texto sea HAM

- Probabilidad de que un texto sea SPAM dado que contiene las palabras W1 a W n :

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

- Probabilidad de cada palabra

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

N w,s es la cantidad de correos SPAM que contienen la palabra W.
N w,h es la cantidad de correos HAM que contiene la palabra W.
N s es la cantidad de correo de SPAM
N h es la cantidad de correos de HAM.

Pruebas de rendimiento

El proceso seguido para hacer las respectivas pruebas de rendimiento fue que se dividió un conjunto de datos en dos partes, una para entrenamiento y otra para pruebas, manteniendo una distribución equitativa de las clases 'spam' y 'ham'. Luego, se calculó la probabilidad de que las palabras en los mensajes de entrenamiento indicarán spam. Con estas probabilidades, se evaluó el modelo en el conjunto de prueba, clasificando cada mensaje como spam o no spam y calculando la probabilidad de cada clasificación. Finalmente, se generaron y mostraron métricas de desempeño para evaluar la precisión del modelo, incluyendo la matriz de confusión y un informe de clasificación. Este fue la matriz que se obtuvo:

```
Matriz de Confusión y Reporte de Clasificación:
[[960   6]
 [ 43 106]]
              precision    recall  f1-score   support

         ham       0.96      0.99      0.98       966
        spam       0.95      0.71      0.81       149

    accuracy                           0.96      1115
   macro avg       0.95      0.85      0.89      1115
weighted avg       0.96      0.96      0.95      1115
```

Figura #6 - Matriz de confusión y reporte de clasificación

En la figura #6, se detalla la matriz de confusión y el informe de clasificación proporcionados, aquí se refleja el rendimiento del modelo de clasificación de mensajes como 'spam' o 'ham'.

- Matriz de Confusión
  - Verdaderos Positivos (ham): 960 casos correctamente identificados como ham.
  - Falsos Positivos: 6 casos erróneamente clasificados como spam.
  - Falsos Negativos: 43 casos erróneamente clasificados como ham.
  - Verdaderos Positivos (spam): 106 casos correctamente identificados como spam.

- Métricas de Desempeño
  - Precisión (Ham): 0.96 indica que el 96% de las predicciones etiquetadas como ham eran correctas.
  - Recall (Ham): 0.99 muestra que el 99% de los mensajes ham reales fueron identificados correctamente.
  - F1-Score (Ham): 0.98, un balance entre precisión y recall, indica un alto rendimiento en la identificación de ham.
  - Precisión (Spam): 0.95 sugiere que el 95% de las predicciones de spam eran correctas.
  - Recall (Spam): 0.71 indica que sólo el 71% de los mensajes de spam reales fueron detectados.
  - F1-Score (Spam): 0.81 refleja un rendimiento aceptable pero mejorable en la identificación de spam.

- Impacto de las Decisiones Tomadas
  Limpieza de Datos:
  Eliminación de palabras vacías y reducción de palabras a su raíz (stemming y lematización): Esta decisión ayudó a concentrar el análisis en las palabras con mayor carga semántica, mejorando la precisión general del modelo.

Normalización a minúsculas y eliminación de caracteres no alfabéticos: Estos pasos garantizan uniformidad en el procesamiento del texto, lo cual es positivo para la consistencia de los datos.

- Cálculos del Modelo:
  Uso del Teorema de Bayes para estimar probabilidades: Permitió incorporar un enfoque probabilístico riguroso basado en la evidencia del entrenamiento, lo cual es fundamental para la alta precisión observada.
  Evaluación basada en umbral para clasificación: La selección de un umbral de 0.5 ha demostrado ser efectiva para ham pero quizás no óptima para spam, dado el menor recall.

- Impacto General
  El proceso de limpieza y los cálculos probabilísticos han llevado a un modelo altamente preciso y eficiente para detectar mensajes ham, con un excelente equilibrio de precisión y recall. No obstante, el rendimiento en la detección de spam, aunque bueno, muestra margen de mejora, especialmente en el recall. Esto sugiere que futuras iteraciones del modelo podrían beneficiarse de una revisión del umbral de decisión y posiblemente de una estrategia más sofisticada para manejar la independencia de características y la diversidad de formas de spam.

Conclusiones

Los resultados obtenidos de la matriz de confusión y el reporte de clasificación ilustran un rendimiento relativamente preciso del modelo en la identificación de mensajes ham, con una precisión y recall notables, evidenciados por un F1-score de 0.98. En contraste, la detección de mensajes spam, aunque con un buen porcentaje, revela que no es tan precisa como con los mensajes ham, particularmente en términos de recall, donde se observa una tasa más baja de 0.71. Este desbalance sugiere que, mientras el modelo es altamente efectivo en minimizar los falsos positivos, tiende a clasificar un número significativo de mensajes spam como ham. Además, las decisiones tomadas durante la fase de limpieza de datos, aunque es notable que fueron uno de los motivos por los cuales la precisión alcanzada fue grande, podrían estar contribuyendo a la pérdida de cierta información útil para la identificación de spam.