Universidad del Valle de Guatemala
Inteligencia Artificial
Sección 30

Proyecto 02

Redes Bayesianas

Brian Carillo 21108

Guatemala, 06 de mayo del 2024

Análisis De Datos Exploratorio (EDA)

En esta fase se definieron características estadísticas relevantes del conjunto de datos proporcionado, con el objetivo de diferenciar los grupos de spam/ham y obtener conclusiones empíricas al respecto.

Descripción del Dataset:

Columna v1: Etiqueta de tipo de mensaje, clasificados en “ham” como textos legítimos y “spam” como contenido no deseado.

Columna v2: Valor de texto del mensaje.

Gráficas:

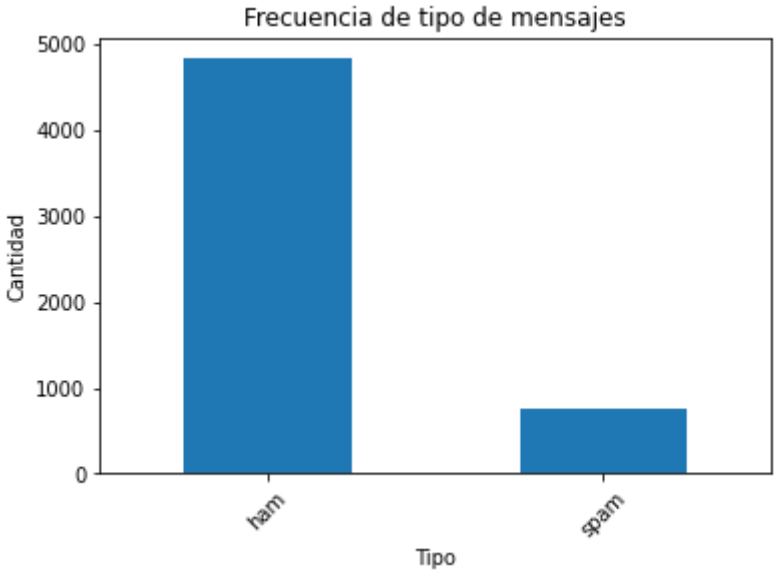

Figura 1. Frecuencia de ham y spam en dataset.

La Figura 1 muestra que existe mayor cantidad de mensajes tipo ham que mensajes tipo spam. La diferencia en la cantidad sugiere que cualquier sistema de filtrado o clasificación debe ser muy preciso para evitar falsos positivos, es decir, clasificar mensajes legítimos como spam, lo cual podría ser perjudicial para los usuarios.

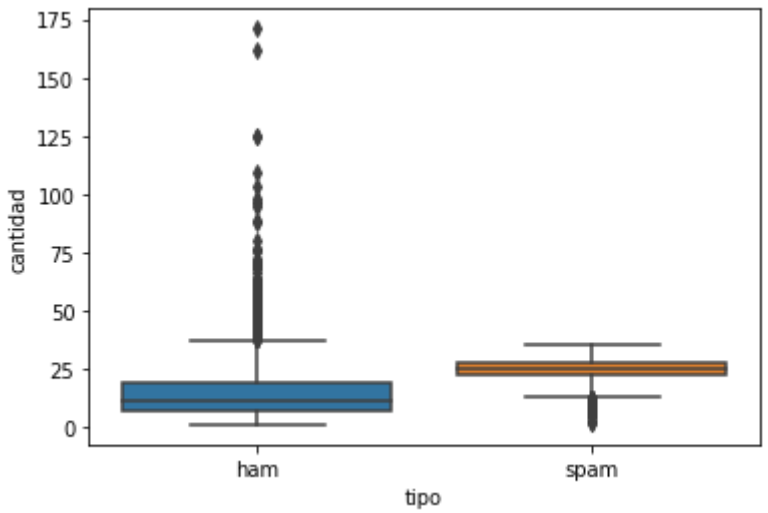

Figura 2. Distribución de la longitud de los mensajes por tipo.

La Figura 2 muestra que en promedio, los mensajes tipo spam son más largos que los mensajes tipo ham. Los mensajes tipo ham presentan mayor dispersión respecto a la media, por lo que las longitudes de los mensajes pueden variar mucho más. Los mensajes 'ham' tienen muchos más valores atípicos que los 'spam', lo que podría indicar casos especiales o extremos en el contenido o las características de estos mensajes.

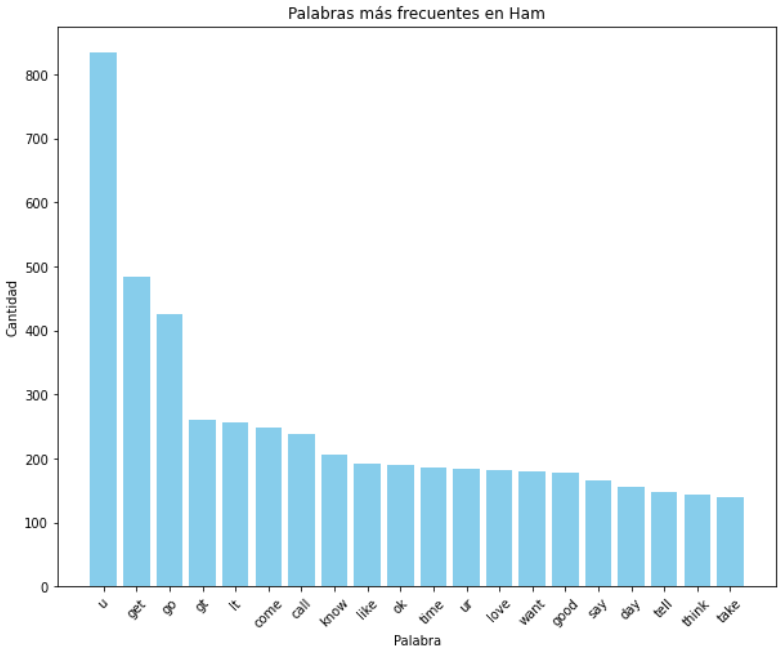

Figura 3. Palabras más frecuentes en Ham.

La Figura 3 muestra que los mensajes ham poseen en su mayoría pronombres y preposiciones, verbos comunes, palabras informales y de cortesía (como like, ok, thanks, love, good), indicadores de contenido personal e informal, etc.

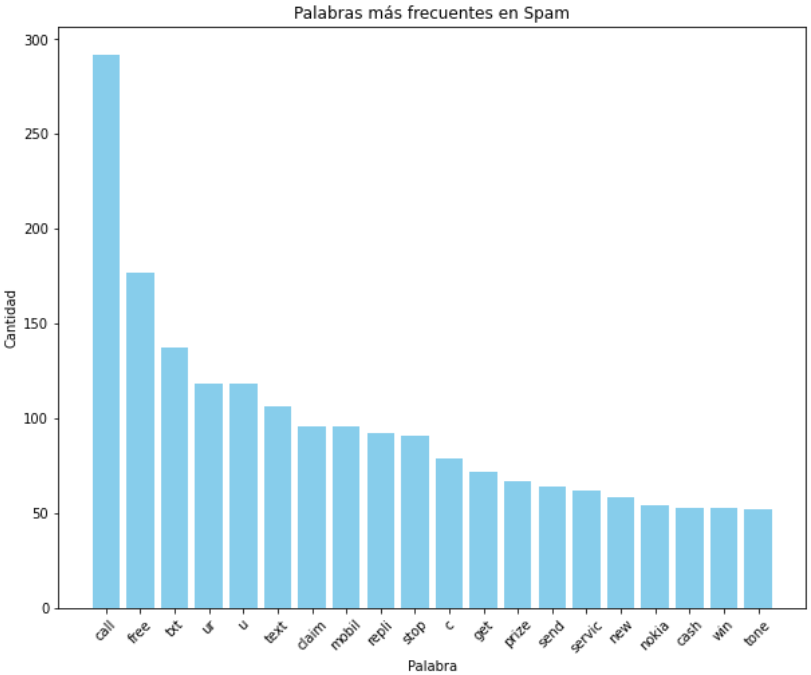

Figura 4. Palabras más frecuentes en Spam.

La Figura 4 muestra que los mensajes spam poseen en su mayoría palabras comunes en ofertas, promociones o solicitudes directas. Así también, se ve una amplia cantidad de llamadas a la acción, abreviaturas, lenguaje informal, mensajes incentivos, etc.

Limpieza de datos

En esta fase se definieron funciones con el objetivo de estandarizar las palabras y filtrar aquellas con un valor relevante para el modelo a generar.

- Conversión a letras minúsculas con el objetivo de realizar un modelo case insensitive.
- Tokenización del texto utilizando word_tokenize, con el objetivo de separar las palabras de los textos.
- Eliminación de tokens con signos de puntuación.
- Eliminación de tokens con números.
- Eliminación de tokens clasificados como stopwords, es decir, palabras con poco valor para el modelo (preposiciones, artículos, etc.).
- Estandarización de las palabras a través del algoritmo de Lemmatization. Esto con el objetivo de llevar las palabras a su forma base, realizando un análisis morfológico sobre cada palabra.
- Estandarización de las palabras a través del algoritmo de Stemming. Puesto que este únicamente corta las terminaciones de las palabras, se utilizó posteriormente a la Lemmatization.

Modelo

En esta fase se definieron las probabilidades necesarias en base al conjunto de entrenamiento, con el objetivo de generar un modelo capaz de predecir si un texto es spam o ham a través de una red bayesiana.

Construcción de conjunto de entrenamiento y prueba:
Se utilizó la librería scikitlearn para poder separar el dataset en un conjunto de entrenamiento y prueba, manteniendo la misma proporción de ham y spam original en ambos sets.

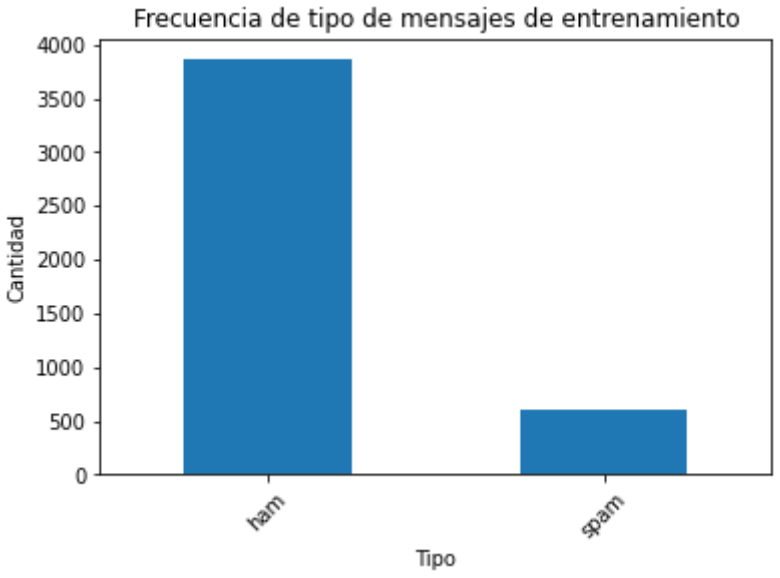

Figura 5. Frecuencia de tipo de mensajes de entrenamiento.

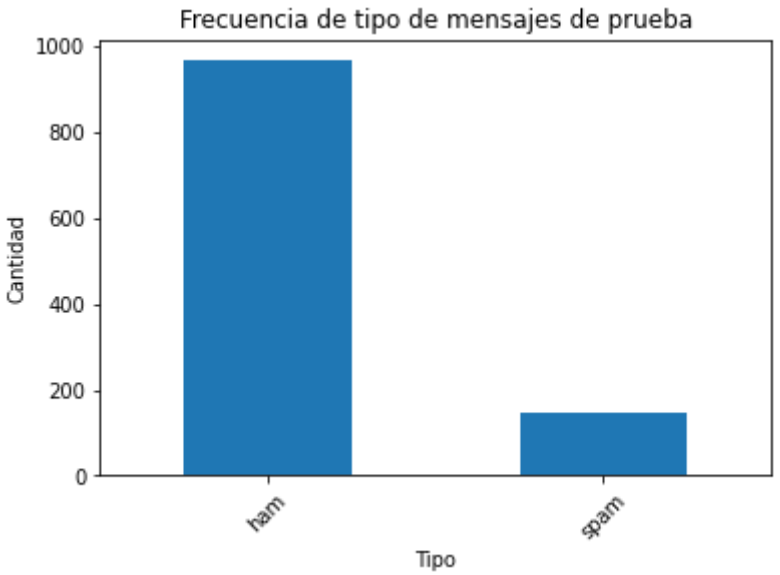

Figura 6. Frecuencia de tipo de mensajes de prueba.

Construcción de diccionarios:
Estos diccionarios contienen los conteos necesarios para obtener las probabilidades que el modelo utilizará posteriormente.

Probabilidades:
Función de probabilidad de Spam dado una palabra específica P(S|W).
Esta función es una implementación de la siguiente fórmula

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

Función de probabilidad de Spam dado un conjunto de palabras W1 a Wn P(S|W).
Esta función es una implementación de la siguiente fórmula

$$P(S|W) = \frac{P_1P_2...P_n}{P_1P_2...P_n+(1-P_1)(1-P_2)...(1-P_n)}$$

Esta última depende de la función de probabilidad de spam dado una palabra específica.

Filtrado de palabras:
Función para reconocimiento de palabras en diccionario de spam o en ham según el conjunto de entrenamiento.

Función para realizar predicción sobre texto:
Esta última función es utilizada para procesar un texto como entrada, y retornar el valor de predicción spam o ham realizada.

Pruebas de rendimiento

En esta fase se realizaron pruebas de predicción sobre el conjunto de prueba definido anteriormente.
Función de Prueba de Predicción:
Tiene por parámetros la categoría del texto y el texto como tal.
Resultados:

```
Matriz de Confusion:
[[102  46]
 [  2 961]]
```

Figura 7. Matriz de confusión.

- El modelo identificó correctamente 102 textos como spam y 961 textos como ham.
- El modelo identificó incorrectamente 46 textos spam como ham y 2 textos ham como spam.

```
Reporte de Clasificacion:
              precision    recall  f1-score   support

         ham       0.95      1.00      0.98       963
        spam       0.98      0.69      0.81       148

    accuracy                           0.96      1111
   macro avg       0.97      0.84      0.89      1111
weighted avg       0.96      0.96      0.95      1111
```

Figura 8. Reporte de Clasificación.

Considerando como positivo el valor de spam, la precisión es de 0.98, el recall de 0.69 y el f1-score de 0.81.

- 0.98 es la proporción de predicciones spam que son correctas, es decir verdaderos positivos entre todos las predicciones positivas.
- 0.69 es la proporción de predicciones spam que fueron capturadas, es decir verdaderos positivos entre cantidad de spam.
- 0.81 es el balance entre precision y recall obtenidos.
- 0.96 es la proporción de predicciones del modelo fueron correctas.

Discusión y conclusiones

El modelo posee una precisión muy alta al identificar spam, sin embargo, posee un valor de recall bajo, lo cual indica que el modelo no logra identificar una cantidad significativa de spam. Esto se explica según el alto valor de Falsos Negativos, es decir textos que el modelo identifica como ham, pero que en realidad son spam.

- Impacto de la limpieza de datos: La estandarización y limpieza de datos permitió que el modelo pueda generalizar los resultados de entrenamiento al conjunto de prueba, por lo que es posible inducir que el alto accuracy se debe al procesamiento previo.

- Impacto de los cálculos del modelo: En el modelo se definió una probabilidad de cero cuando el denominador sea igual a este, como una medida preventiva a la zero division. Esto repercute en casos en los que la combinación de las probabilidades de diferentes palabras generan que el término [P(W|S) * P(S)] + [P(W|H) * P(H)] sea igual a cero. Por lo tanto, es posible inducir que muchos de los textos de prueba fueron clasificados como ham ante esta medida previamente descrita.

- Conclusiones a nivel general: El modelo es bastante robusto tanto en identificar spam como en evitar marcar incorrectamente los textos ham. Sin embargo, la existencia de falsos positivos sugiere que podría ser útil revisar y posiblemente ajustar los criterios y características utilizadas por el modelo, para reducir aún más el número de textos legítimos marcados como spam.