Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

Sistema en Python que detecta SPAM

Escenario PBL1

Análisis de métodos de razonamiento e incertidumbre

José Miguel Pérez Flores A00832401

Kevin Montoya Campaña A01740352

José Pablo Sánchez González A01412539

Ricardo de Jesús Balam Ek A00831262

Carlos Mateos Pérez A01654085

Supervisado por
Marco Otilio Peña Díaz

Monterrey, Nuevo León. Fecha, September 26, 2022

1 Problematización

SPAM es el término usado para referirse a los e-mails no solicitados, es cualquier mensaje enviado a varios destinatarios que no solitaron especificamente tal mensaje. Detectar estos mensajes es una prioridad para que mensajes importantes no queden perdidos entre tanta basura informática (ESET, 2022). Los mensajes de SPAM acumularon el 45.37% del tráfico de e-mails enviados en diciembre de 2021. Durante el periodo medido más reciente, Rusia generó la mayor cantidad de e-mails de SPAM con 24.77% del volúmen global de SPAM (Statista, 2022). Es casi imposible pensar en los e-mails sin relacionarlos de alguna manera con el SPAM, en 2021, 319.6 billones de e-mails fueron enviados y recibidos en una frecuencia diaria.

2 Enfoque

El trabajo consiste en realizar un modelo tal que trabaje como un clasificador de correos y los pueda diferenciar entre uno normal y uno SPAM. Será realizado con el clasificador naive bayes. Buscamos que este clasificador sea lo mas simple posible.

3 Propósito

Implementar un algoritmo que haga una interfaz de correos electrónicos amigable, con la capacidad de seccionar los correos no deseados en una categoría aparte de la bandeja principal.

4 Información

Un clasificador Naive Bayes es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. En términos simples, asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable (Gandhi, 2018). Por ejemplo, un mensaje SPAM puede tener tendencias a palabras como: GRATIS, FELICIDADES o INGRESE. Un clasificador de Naive Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que el mensaje sea SPAM, independientemente de la presencia o ausencia de las otras características. Se tendrá una base de datos de 5559 ejemplos que además detalla si el mensaje es uno normal o SPAM. Esto ayudará a clasificar las probabilidades de las palabras a cada categoría.

5 Razonamiento

Se comenzó con una planeación de cómo se debería de acomodar los datos de tal manera que puedan procesarlos un modelo base. Se comenzó haciendo un Wrangling de los datos. En otras palabras, se eliminaron ciertas palabras y símbolos de los mensajes de tal manera que sólo quedaran las características relevantes para el modelo. Se simplificaron grupos de palabras a sus fonemas base, como las distintas conjugaciones de verbos, entre otros. Esto mejora mucho al modelo al reducir el numero de palabras existenes, aumentar su frecuencia en los mensajes y por lo tanto reducir la variabilidad. Ademas, solamente trabajamos con las 1500 palabras mas usadas en nuestra base de datos, para reducir el impacto de palabras poco comunes que podrian causar ruido en los datos. Para terminar se clasificó las condiciones de SPAM o mensaje normal con un solo bit alto o bajo (0 y 1).

Nuestra prioridad fue mantener la simpleza del modelo naive bayes, por lo que no se utilizaron técnicas mas avanzadas como uso de TF-ICF para determinar la importancia relativa de cada palabra. Consideramos que representar la utilidad del Naive Bayes en su forma mas simple era lo mas importante del proyecto, por lo que se realizo el menor preprocesamiento posible.

Después se comenzó a trabajar en el algoritmo. Mediante las librerías que seleccionamos hicimos una clase base que en cada función que manejaba, desarrollaba una diferente tarea. Una calculaba las probabilidad de que una palabra aparezca en un mensaje de texto dado que es spam y otra dado que no es SPAM. Después, junto con estas probabilidades se formó la función que calculaba la probabilidad total, es decir, palabra por

palabra de cada mensaje para así obtener una comparativa de cuál sería la mayor probabilidad de que sea SPAM o no. Finamlente, se resuelve todo en una función que acomoda ambas probabilidades y forma un booleano que muestra un resultado verdadero o falso, dependiendo de si los niveles de SPAM sean mayores a los que no. En otras palabras, muestra un resultado verdadero para SPAM y uno falso para un mensaje normal. Estas funciones fueron aplicadas a todo la base de datos de prueba para obtener que tan eficaz es nuestro modelo, pero tambien se puede aplicar a cualquier mensaje en inglés que queramos probar.

6 Conclusiones

Para concluir el modelo, se obtuvieron valores que corroboran la exactitud y precisión del modelo, de igual manera, mediante una matriz de confusión, se proyectaron los resultados, tanto los verdaderos como los falsos. Finalmente, se realizaron pruebas para corroborar el trabajo. A partir de la matriz de confusión (1) se obtuvieron las siguientes métricas, el programa presenta una precisión de 61.4%, un F1 score de 71.8% y una exactitud de 92%.

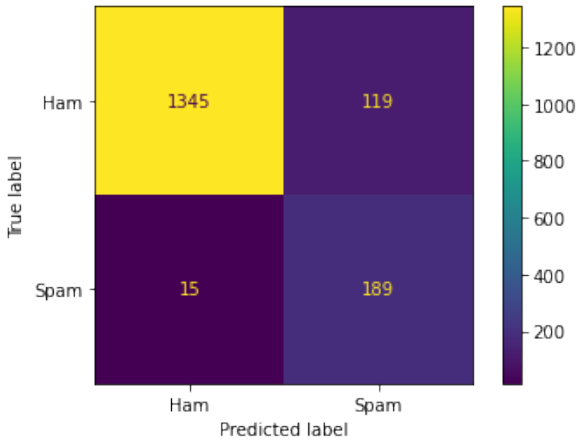

Figure 1: Matriz de confusión

Finalmente, el modelo funciona correctamente ya que se examina con distintas frases para conocer si logra reconocer entre mensajes de SPAM y mensajes normales. (2)

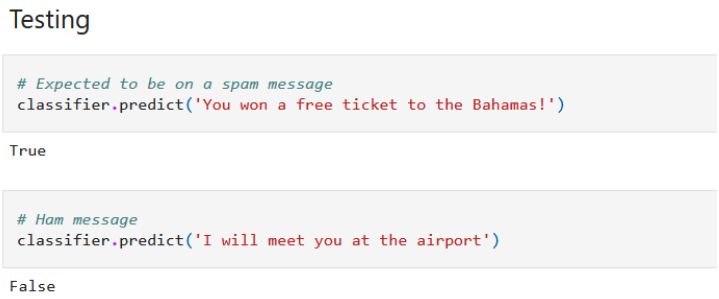

Figure 2: Testing del modelo

Gracias a estas metricas y los ejemplos probados determinamos que el modelo de Bayes Naive si es apropiado para aplicar a la clasificación de emails como spam y no spam. Aunque podrian existir modelos mas precisos, este es extremadamente simple y se puede aplicar en distintos ambitos.

References

ESET. (2022). ¿qué es el spam?

Gandhi, R. (2018). Naive bayes classifier. *Towards Data Science*. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Statista. (2022). Global spam volume as a percentage of total e-mail traffic from january 2014 to december 2021, by month.