Detección de Spam en Email

Jose Adrian Castillo Sierra

Problemática

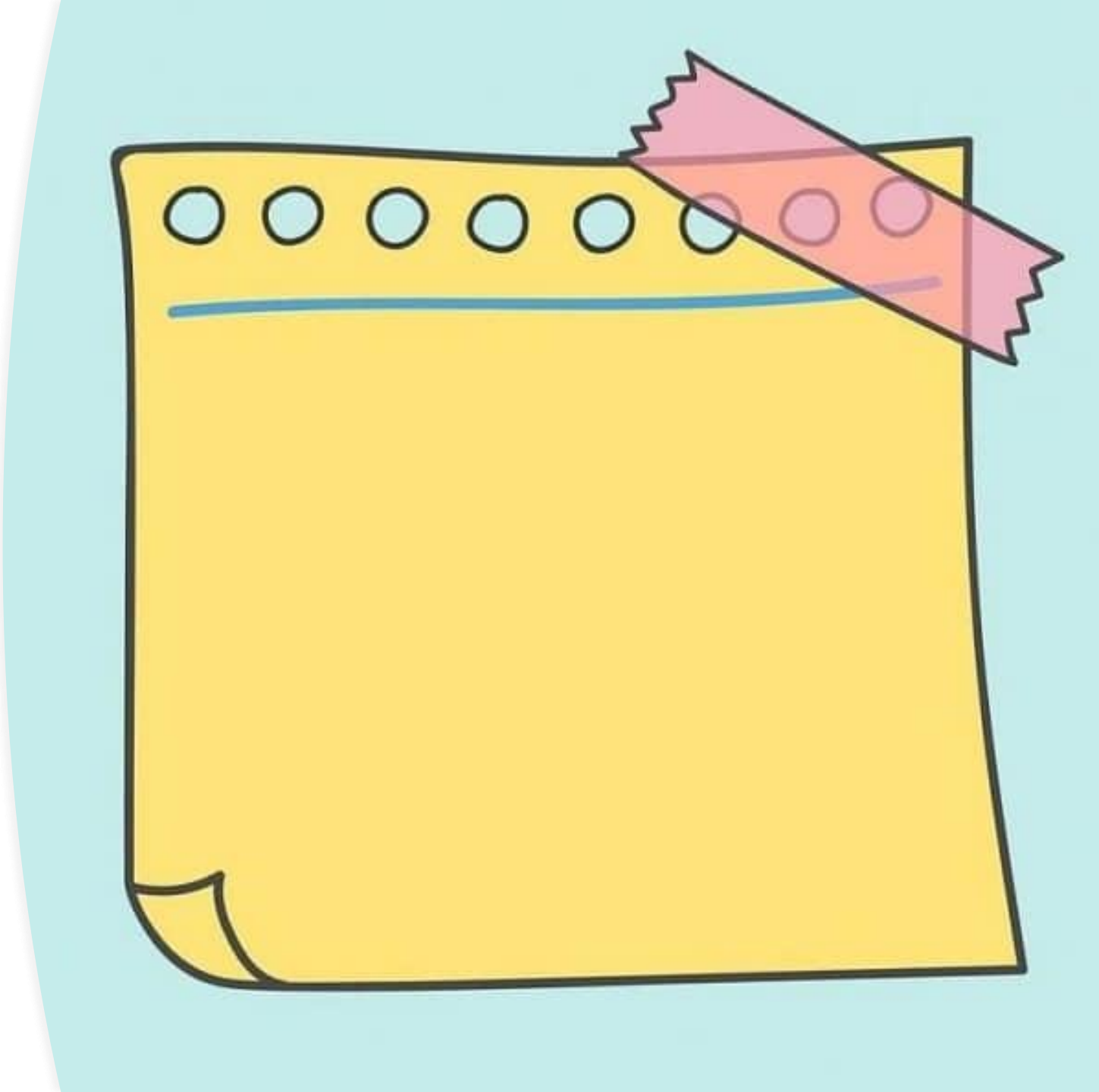
La atención a cliente es una de las partes mas importantes de cualquier negocio. No importa si nuestro producto tiene ventajas sobre la competencia si no podemos atender debidamente a los clientes cuando presenten problemas técnicos nuestro producto no es competitivo.

Problemática

En este proyecto se plantea la clasificación de correos de acuerdo a su contenido para asignar importancia de acuerdo al problema de los usuarios.

Nota

No se pudo utilizar un dataset perteneciente a mi trabajo debido a que en ocasiones los clientes comparten innecesariamente fotos y/o datos de tarjetas bancarias. Por este motivo se utilizo un dataset similar para trabajar este problema.



EDA

Distribución de Clasificaciones

EDA

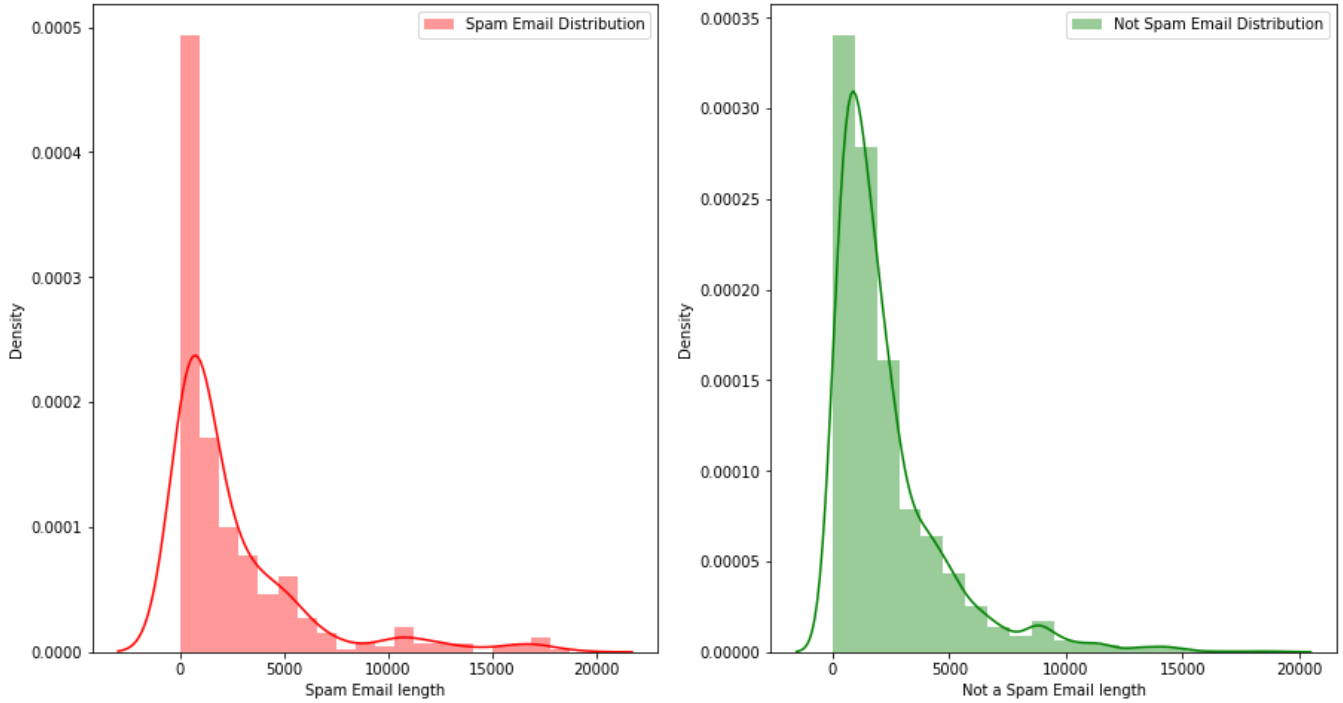
WordCloud Correos Normales

EDA

WordCloud Correos Spam

EDA

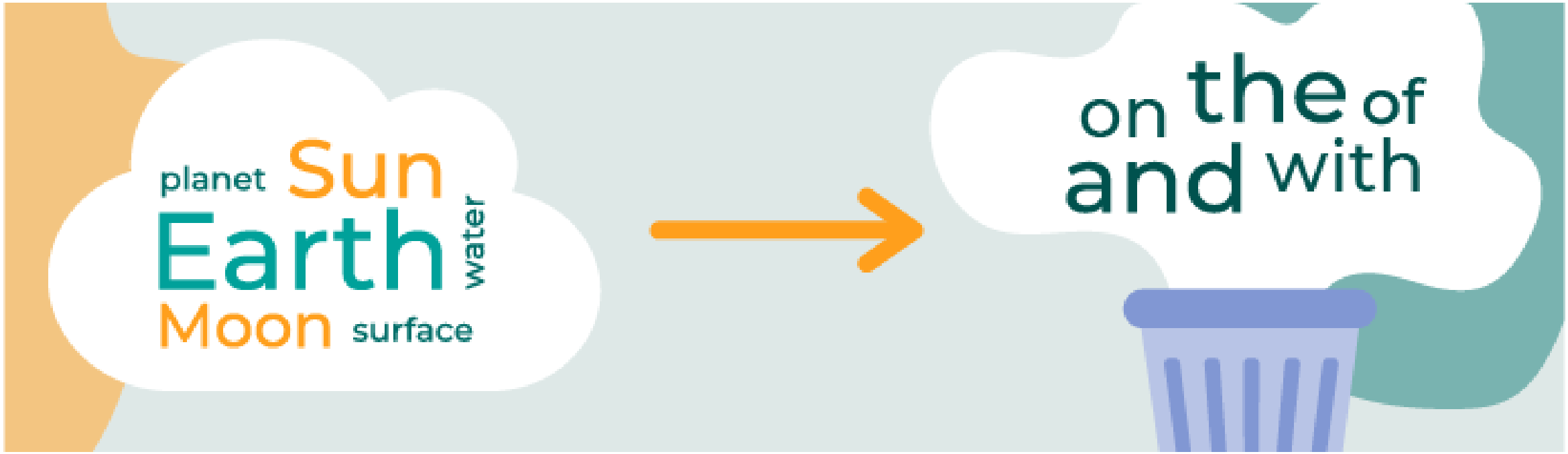
Histograma de Palabras por Correo



Bines de 1000

Preprocesamiento

Eliminacion de stopwords



Preprocesamiento

Sustitución de mails, números de teléfono y maldiciones

<janesmith@email.com>

Calculating TF-IDF (very simple example)

Solution:

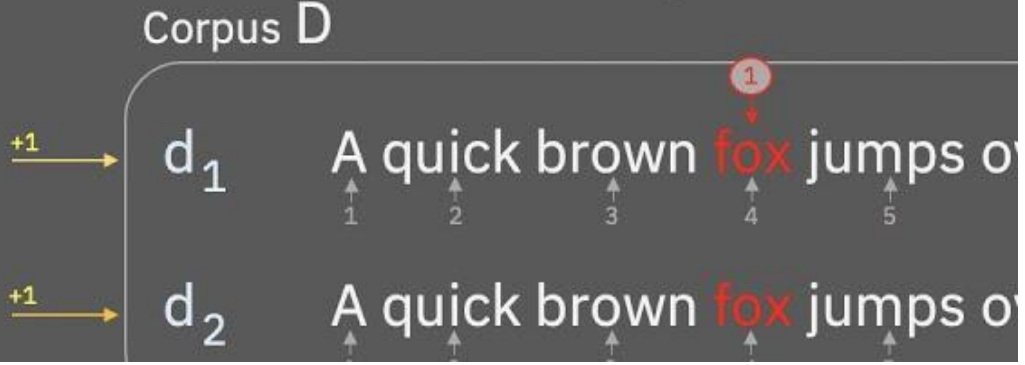
TF-IDF

TF is the frequency of any "term" in a given "document".

IDF is constant per corpus, and accounts for the ratio

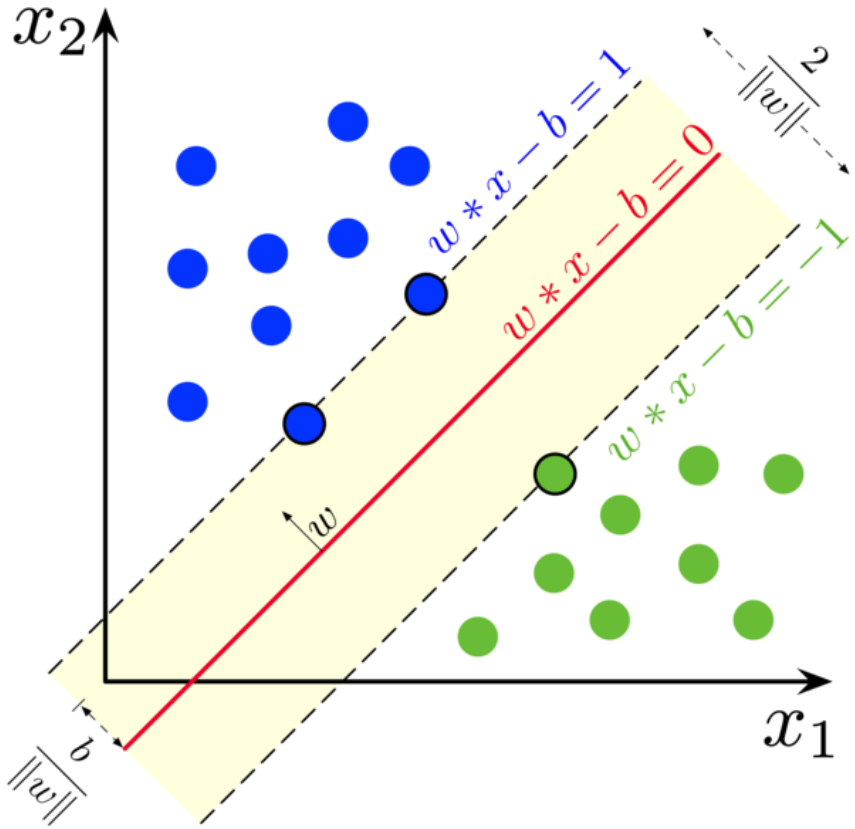
TF("fox", d1) = 2 / 12 = 0.17

TF("fox", d2) = 3 / 12 = 0.25

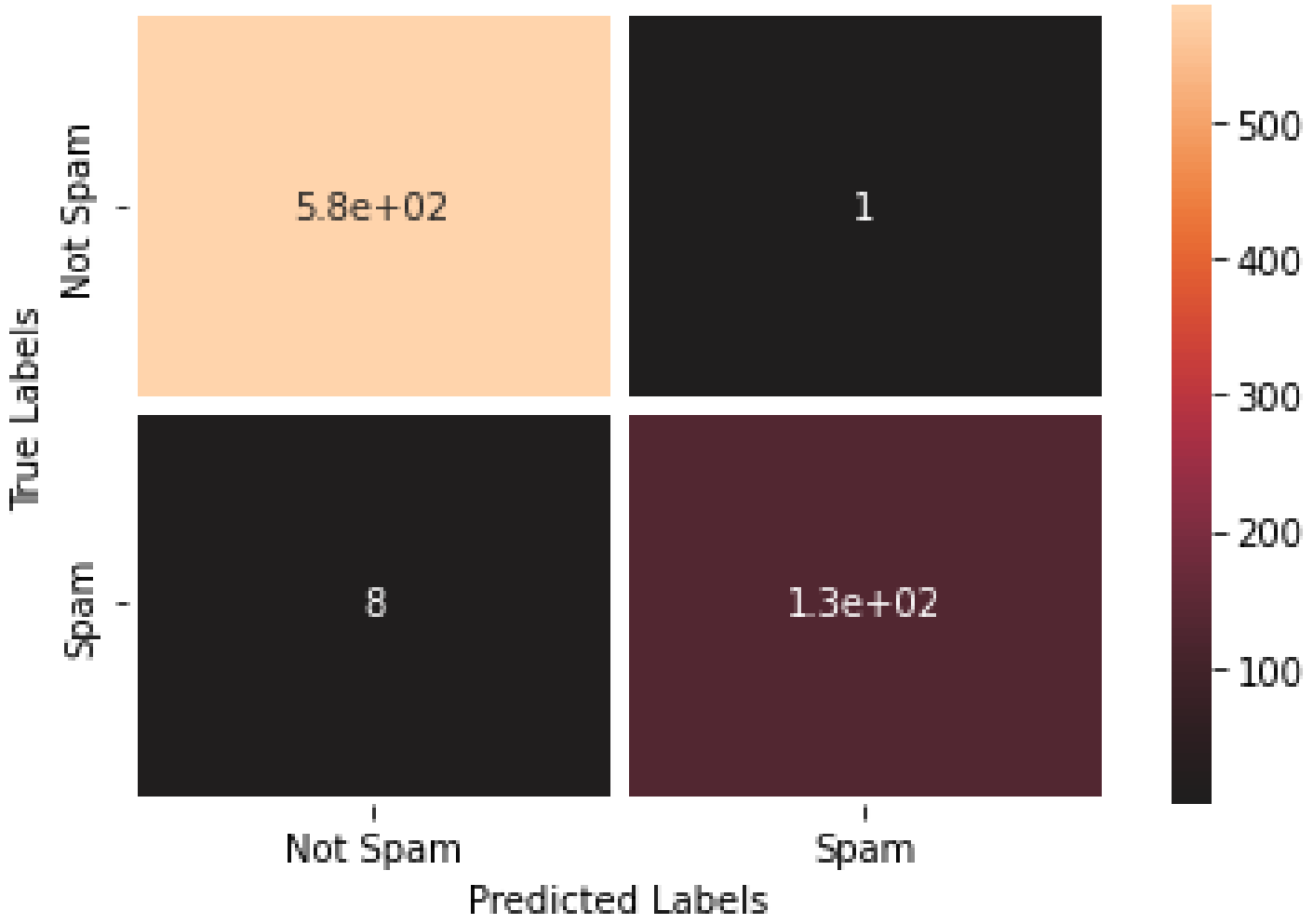


Entrenamiento

Support Vector Machine



Resultados



Accuracy: 98.75%

Próximos Pasos

Aplicar los procesos utilizados en este proyecto para el etiquetado de los correos recibidos por el departamento de *Customer Care*.