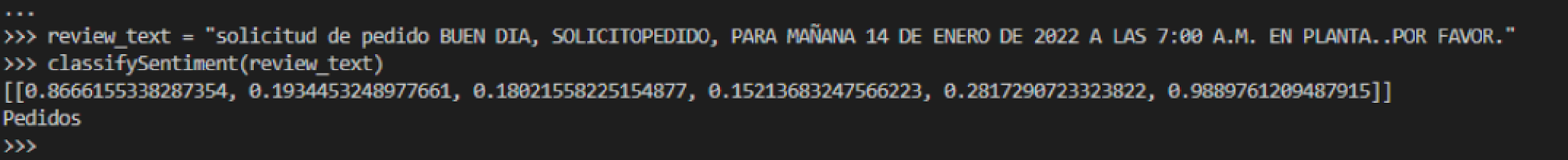
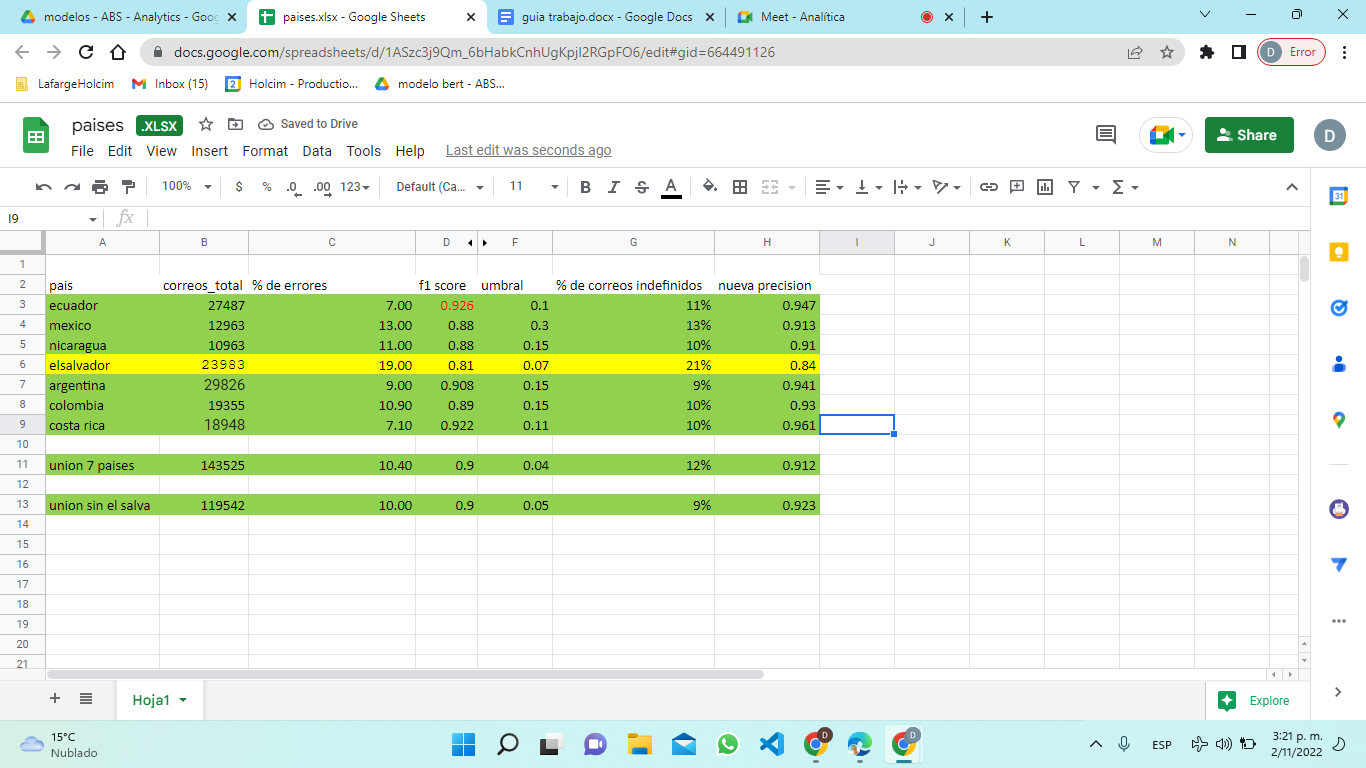
1. Utilizamos la base de cada país para entrenar el modelo bert de cada país (El modelo bert es un modelo pre entrenado en varios idiomas con millones de palabras capaz de entender oraciones y entender el lenguaje como un humano), el modelo se descarga se le agrega una capa neuronal para que se especialice en la solución de nuestro problema.
2. Deberá instalar la librería Transformers en la consola **CMD** (ejecutar como administrador) para poder ejecutar el código de Python, en esta página encontrará la guía de las librerías que necesita instalar:

<https://huggingface.co/docs/transformers/installation>

1. Utilizamos 6 categorías para méxico para clasificar los correos: 'Aplicación de pago y pedido': 0, 'Datos maestro': 1, 'Aplicación de pagos': 2, 'No cliente': 3, 'Otros': 4, 'Pedidos': 5, y siete categorías para los demás países.
2. Decidimos unir las columnas **Breve descripción** y **Descripción,** en una sola columna, y esa nueva columna la utilizamos para entrenar el modelo.
3. Utilizamos 3 ciclos (los datos entran 3 veces al modelo) debido al tiempo que toma hacer esto, entre más ciclos es mejor la clasificación, pero llega un punto en que más ciclos no aumenta la precisión.
4. Dividimos los datos en datos de entrenamiento y validación y por ser datos desbalanceados la métrica es F1.
5. Creamos una función para clasificar nuevos correos, esta función nos arroja la categoría a la que pertenece el nuevo correo, nos arroja las probabilidades de pertenecer a cada categoría, la categoría con la probabilidad más alta es a dónde se clasifica el correo.
6. 
7. Para probar la función de clasificación y para hallar el umbral, utilizamos los correos de validación
8. Identificamos que cuando un correo queda mal clasificado, la segunda probabilidad mas alta, es un valor muy elevado. Por eso establecimos un umbral para cada país, todo correo cuya segunda probabilidad supere este valor será considerado como indefinido, ese correo se clasifica en una nueva categoría llamada “Indefinido”, este correo deberá ser clasificado manualmente. Este límite se calculó teniendo en cuenta que el número de correos indefinidos no fuera demasiado grande, ya que entre menor el límite, mayor es el número de correos indefinidos, y se busco atrapar el mayor número de correos mal clasificados,
9. Por ejemplo para Ecuador concluimos que el mejor umbral era del 10%, quiere decir que los correos cuales segundas probabilidades sean superiores a 0.90, serán clasificados en la categoría “indefinidos”, con este umbral el 11% de los correos en total son clasificados como indefinidos y la precisión es ahora del 0,947.
10. Para poder ejecutar el modelo se necesitan 3 archivos, el script de Python llamado “**CODIGO\_EJECUTAR\_pais**”, el archivo llamado “**finetuned\_BERT\_epoch\_3”nombre\_pais**” y el archivo “**possible\_labels**” el cual contiene las categorías , el archivo llamado **“código\_entrenamiento”** es todo el entreno del modelo el cual no es necesario correrlo, en este archivo esta la funcion para hallar las segundas probabilidades.
11. En estas dos páginas encontrara ejemplos parecidos de cómo desarrollar el modelo: <https://github.com/codificandobits/Analisis_de_sentimientos_con_BERT/blob/master/BERT_analisis_sentimientos.ipynb>

<https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613>