**Forma

Descripción generada automáticamente con confianza media**

**BOOTCAMP:**

IA & BIG DATA

**Detección de reseñas positivas y negativas en una tienda de ropa local**

Juan Manuel Moreno Ruiz

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

Use este documento como base de su TFG (eliminando todo lo que sobre. Observe el texto marcado con fondo amarillo y modifíquelo con la información pertinente). Esta plantilla contiene todos los estilos que deben ser usados. Es muy conveniente que automatice mediante los estilos el documento (Capítulos, Apartados, Subapartados…), de forma que se puedan generar los índices de forma automática.

En las próximas hojas se indican las normas y estilos necesarios, así como algunas recomendaciones para la realización del TFG.

**ÍNDICE**

[1 PROBLEMA A RESOLVER 3](#_Toc170471160)

[1.1 DESCRIPCIÓN DEL PROBLEMA 3](#_Toc170471161)

[2 OBJETIVO 3](#_Toc170471162)

[3 SOLUCIÓN ELEGIDA 4](#_Toc170471163)

[3.1 RoBERTa 4](#_Toc170471164)

[3.2 ¿Por qué elegimos RoBERTa? 4](#_Toc170471165)

[4 PROCEDIMIENTO DE RESOLUCIÓN 4](#_Toc170471166)

[4.1 Elección de dataset 4](#_Toc170471167)

[4.2 Preprocesamiento de datos 5](#_Toc170471168)

[4.3 Entrenamiento y validación del modelo 5](#_Toc170471169)

[4.4 Evaluación del modelo 6](#_Toc170471170)

[5 DESPLIEGUE DE SOLUCIÓN 7](#_Toc170471171)

[5.1 Herramienta de despliegue 7](#_Toc170471172)

[5.2 Enlace al modelo desplegado 7](#_Toc170471173)

# PROBLEMA A RESOLVER

## DESCRIPCIÓN DEL PROBLEMA

El problema principal a resolver es el aumento de devoluciones en una tienda de ropa.

Las devoluciones de productos no solo impactan negativamente en las finanzas de la tienda, sino que también pueden indicar insatisfacción del cliente con los productos recibidos.

Identificar las causas subyacentes de las devoluciones es crucial para implementar estrategias efectivas que puedan reducir esta tasa.

# OBJETIVO

El objetivo principal del proyecto es abordar el problema de las devoluciones en la tienda de ropa. Para ello, se plantean los siguientes objetivos:

* Detectar causas de devoluciones: Utilizar el modelo *XLM-RoBERTa* para analizar reseñas de clientes y detectar las razones por las cuales devuelven la ropa. Esto incluye identificar aspectos como el tamaño incorrecto, problemas de calidad, discrepancias entre la descripción del producto y la realidad percibida, entre otros.
* Aumentar las ventas: Al comprender mejor las necesidades y expectativas de los clientes, se busca mejorar la experiencia de compra y aumentar la satisfacción del cliente, lo cual debería traducirse en un incremento de las ventas.
* Reducir las devoluciones: Implementar estrategias basadas en los resultados obtenidos del análisis de reseñas para reducir la tasa de devoluciones. Esto puede incluir mejoras en la descripción de productos, políticas de devolución más claras, recomendaciones de tamaño más precisas, entre otras medidas.

# SOLUCIÓN ELEGIDA

Para abordar el análisis de reseñas y la detección de reseñas positivas y negativas, se ha elegido el modelo: *XLM-RoBERTa (base-sized model)*, de HuggingFace (<https://huggingface.co/FacebookAI/xlm-roberta-base>).

## RoBERTa

RoBERTa es un modelo avanzado para procesamiento del lenguaje natural desarrollado por Facebook AI Research (FAIR). Está basado en la arquitectura de transformers y es una versión mejorada del modelo BERT de Google. RoBERTa cuenta con una estructura más grande y mejoras en su método de entrenamiento.

El modelo RoBERTa utiliza un enfoque de pre-entrenamiento no supervisado para aprender representaciones de alta calidad del lenguaje natural. Esto se logra mediante una técnica en la que el modelo debe completar oraciones con palabras faltantes en el corpus de entrenamiento.

Una vez pre-entrenado, RoBERTa puede aplicarse a diversas tareas de procesamiento de lenguaje natural, como la clasificación de emociones, la generación de texto y la traducción de idiomas. RoBERTa ha mostrado un rendimiento muy adecuado en muchas de estas tareas, convirtiéndose en uno de los modelos pre-entrenados más utilizados actualmente en el campo del lenguaje natural.

## ¿Por qué elegimos RoBERTa?

Este modelo se selecciona debido a su capacidad para entender el contexto y las sutilezas del lenguaje, especialmente en textos en español, como en nuestro caso, permitiendo una clasificación más precisa de reseñas positivas y negativas.

Esta elección se prefiere sobre modelos como *DistilBERT base model (uncased)*, debido a la robustez y precisión demostrada en tareas de análisis de sentimientos, cruciales para identificar las opiniones de los clientes respecto a los productos de la tienda de ropa. Además, DistilBERT, fue entrenado principalmente en textos en inglés.

Por otro lado, se eligió el modelo BERT y no otro modelo (como LLM) por:

* Arquitectura bidireccional
* Eficiencia en tareas específicas de clasificación
* Facilidad de implementación

# PROCEDIMIENTO DE RESOLUCIÓN

## Elección de dataset

El dataset elegido es el siguiente: <https://www.kaggle.com/datasets/shavilyarajput/clothing-shoes-and-jewellery-reviews>

Se eligió este dataset por las siguientes razones:

* Variedad de reseñas
* Distribución de reseñas equilibradas
* Formato adecuado para el entrenamiento del modelo

El dataset contiene las siguientes columnas:

* **overall:** calificación de reseña.
* **reviewText:** texto de la reseña en cuestión.
* **summay:** resumen de la reseña.

El total de líneas es de: 792000.

## Preprocesamiento de datos

1. Carga de dataset
2. Traducción del dataset original a español, usando la librería: ***easymnt***
3. División del dataset en conjuntos de entrenamiento, validación y test.
4. Se guarda cada subconjunto en archivos CSV: train.csv, validation.csv y test.csv
5. Se cargan los CSV en DataFrames de pandas
6. Se filtran los ejemplos cuya calificación sea igual a 3, ya que se tratan como nuetras, por lo que se excluyen.
7. Polarizar o categorizar las variables de calificaciones de la siguiente manera:
   * Si calificación = 1 o 2, se sustituye por un 0
   * Si calificación es 4 o 5, se sustituye por un 1
8. Tokenización del texto, usando el modelo: ‘xlm-roberta-base’

## Entrenamiento y validación del modelo

Para el entramiento y validación del modelo, hemos usado la métrica: *Accuracy* y *Loss,* para obtener una medida del rendimiento del modelo.

Como podemos ver en la siguiente captura, en cada fase de entrenamiento obtenemos unas métricas bastante buenas, observando un valor de Accuracy mayor al 90% en la fase de Entrenamiento y Validación.

El proceso de entrenamiento y validación se explica con más detalle en el Notebook.

Imagen que contiene Escala de tiempo

Descripción generada automáticamente

Por tanto, existe un gran rendimiento del modelo.

Aquí podemos observar un gráfico representativo de las métricas mencionadas anteriormente:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

## Evaluación del modelo

En la evaluación del modelo, se vuelve a observar un rendimiento muy bueno en el conjunto de Test:

Texto

Descripción generada automáticamente

# DESPLIEGUE DE SOLUCIÓN

## Herramienta de despliegue

Para el despliegue de la aplicación desarrollada para abordar el problema de aumento de devoluciones en una tienda de ropa, se utilizó ***Streamlit***, una herramienta de código abierto que facilita la creación y el despliegue de aplicaciones web interactivas para aprendizaje automático y análisis de datos.

Streamlit permite transformar modelos y scripts de Python en aplicaciones interactivas con una interfaz de usuario intuitiva, donde los usuarios pueden ingresar reseñas de productos y obtener instantáneamente el análisis de éstas en positivas y negativas utilizando el modelo ROBERTA XLM entrenado.

El despliegue se realiza de manera sencilla ejecutando un único comando en la terminal, lo que hace accesible la aplicación a través de un navegador web. Esto asegura que tanto los stakeholders internos como externos puedan acceder fácilmente a la herramienta y utilizarla para comprender mejor las opiniones de los clientes y tomar decisiones informadas para reducir las devoluciones y mejorar la satisfacción del cliente

## Enlace al modelo desplegado

A continuación, se adjunta el enlace al modelo desplegado para su uso:

* Enlace al modelo desplegado: <https://reviewsiabigdata.streamlit.app/>

**Bibliografía**

* *Dataset:* [*https://www.kaggle.com/datasets/shavilyarajput/clothing-shoes-and-jewellery-reviews*](https://www.kaggle.com/datasets/shavilyarajput/clothing-shoes-and-jewellery-reviews)
* *Modelo:* [*https://huggingface.co/FacebookAI/xlm-roberta-base*](https://huggingface.co/FacebookAI/xlm-roberta-base)