



# Generación de texto condicionado

Matías Lombardi - Patrick Dey

Tutor: Francisco Pérez Sammartino



# Tabla de contenidos

01

Introducción

02

*Datasets*

03

Arquitectura

04

Entrenamiento

05

Modelos  
finales

06

Conclusión



# 01 Introducción



# Problema

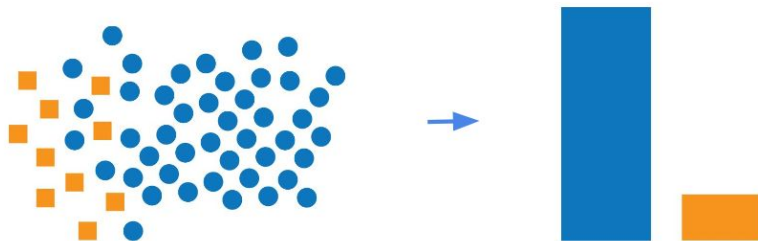
Una de las dificultades más grandes en el ámbito de *Machine Learning* es la **adquisición de datos** debido a:

- Información **incompleta**
- Datos que requieren de **intervención humana** (para etiquetarlos, por ejemplo)
- Conjunto **desbalanceado**
- Entre otras...

Probablemente, la **capacidad generalizadora** de los modelos se vea afectada por estos ítems.

# Propuesta

- Analizar la capacidad generadora de los modelos que constituyen el **estado del arte** en cuanto al análisis de texto.
- Generar datos que **conserve** la **distribución original**.
- Evaluar el **rendimiento** de estos modelos.





# 02

# ***Datasets***

*Corpus* utilizados para el entrenamiento

**Se aplicará el análisis  
de modelos al  
contexto de las  
reseñas de películas**





# Condiciones ideales

- *Dataset* lo suficientemente grande para **entrenar** los modelos
- Escala de **sistema de puntajes** común para todas las entradas
- Contenido **específico** de las películas
- Entradas en el mismo idioma (en este caso **inglés**)





# Analizados

## Amazon

- 8 millones de entradas.
- Películas compradas en la plataforma.
- Reseña habla más sobre el producto que de la película en sí.

• ~~DESCARTADO~~

I have all of the doo wop DVD's and this one is as good or better than the 1st ones. Remember once these performers are gone, we'll never get to see them again. Rhino did an excellent job and if you like or love doo wop and Rock n Roll you'll LOVE this DVD !!

## IMDB

- Cercano a 1 millón de entradas.
- La mayoría de las reseñas refieren exclusivamente del contenido de la película.
- Puntajes del 1 a 10.

This movie is full of suspense. It makes you guess about what is real and what is not. It happens more than once that you have to wonder about what is the truth and who is lying. Because you are just as clueless as the main character, Michelle, you really get to experience the same type of emotions and confusion as she is.

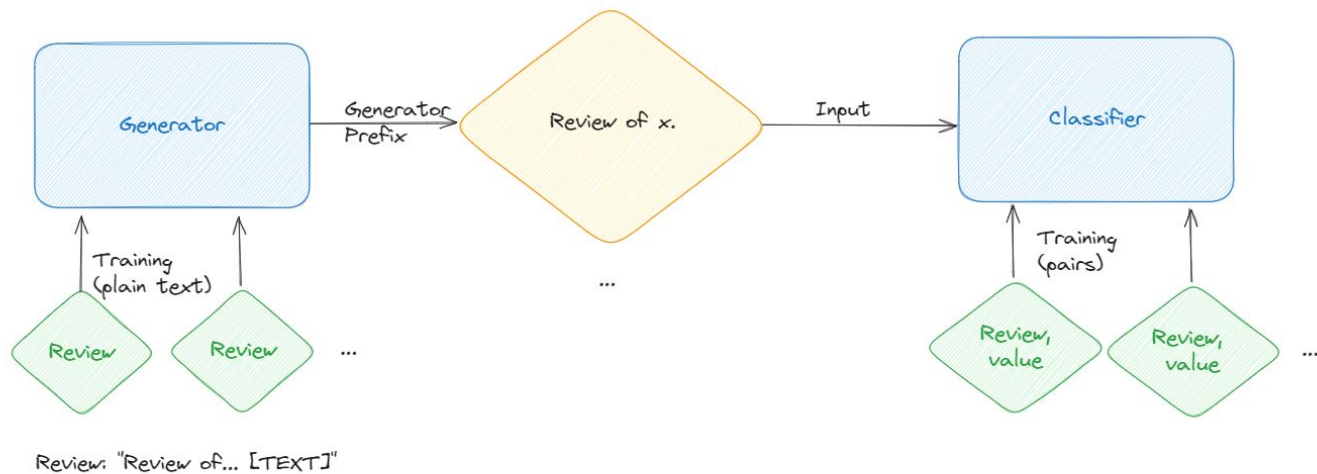


03

# Arquitectura

Modelos y técnicas utilizadas

# Arquitectura final





# Transformers

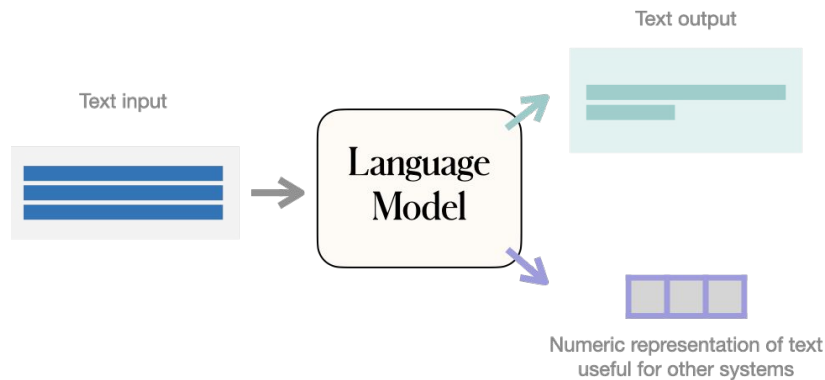
- Los **transformers** constituyen el [estado del arte](#) en cuanto al procesamiento del lenguaje natural.
- La plataforma **HuggingFace** permite utilizar y [ajustar parámetros](#) de modelos entrenados, por lo que se realiza *fine-tuning*.



HUGGING FACE



# Red Generadora



## GPT-2

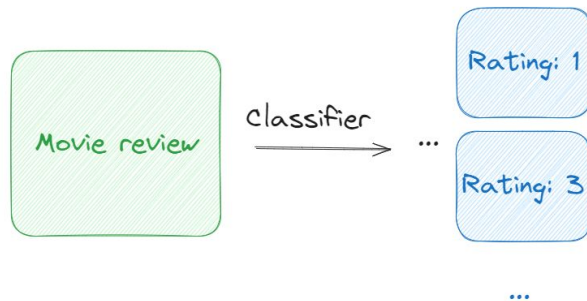
*Generative Pre-trained Transformer*

*Versión distil*

## Único modelo

## Generación sobre un prefijo

# Red Clasificadora



## BERT

*Bidirectional Encoder Representations  
from Transformers*

*Versión distil*

## Reseñas etiquetadas

## Clasificación en 5 clases

# Congelamiento de capas



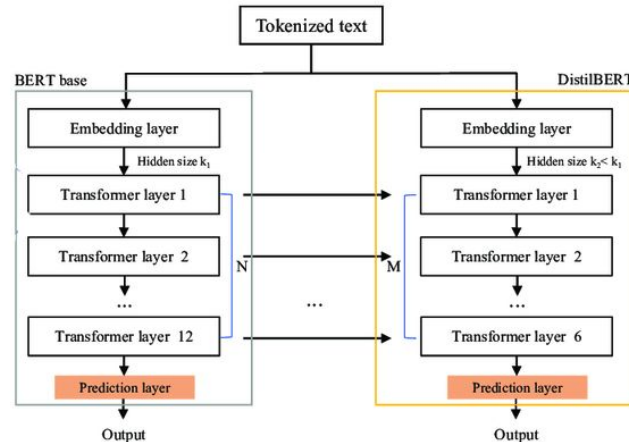
Los modelos consisten de 6 bloques de Transformers y la capa específica a la tarea



Se entrena únicamente el último bloque y la capa final



El resto de las capas quedan fijas



# Preprocesamiento

Para ejemplificar las etapas del preprocesamiento realizado sobre el dataset, se utiliza la siguiente oración para realizar un seguimiento:

Something like this movie won't happen again. This is literally the best superhero movie ever!!!  
<br/><br/>Thank u MCU 😊"











# Preprocesamiento

1. Limpieza del *dataset* original
  - Borrado de
    - i. Emojis
    - ii. Hipervínculos
    - iii. Elementos HTML
    - iv. Símbolos que no son letras, números, comillas, signos de exclamación, puntos, paréntesis, comas
  - Borrado de entradas que contienen insultos (representadas por \*) o puntaje faltante.
  - Recorte de espacios, signos de puntuación y exclamación.

Something like this movie won't happen again. This is  literally the best superhero movie ever!!!   Thank u  
MCU 



# Preprocesamiento

2. Transformación de puntajes a 5 etiquetas posibles:

$$\text{rating} = \left\lceil \frac{\text{rating}}{2} \right\rceil$$

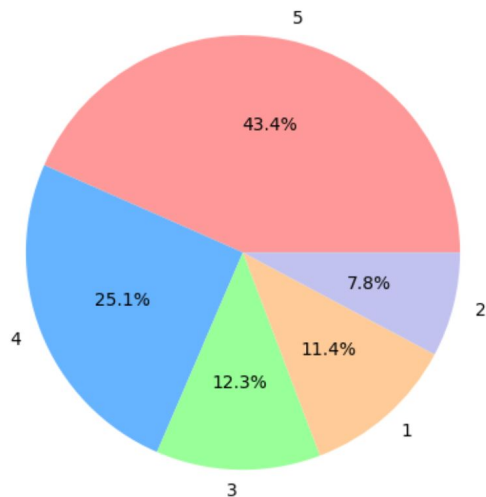
3. Agregado del prefijo a las reseñas utilizadas por el generador:

**Review of X.** This movie is full of suspense.It makes you guess  
about what is real and what is not...

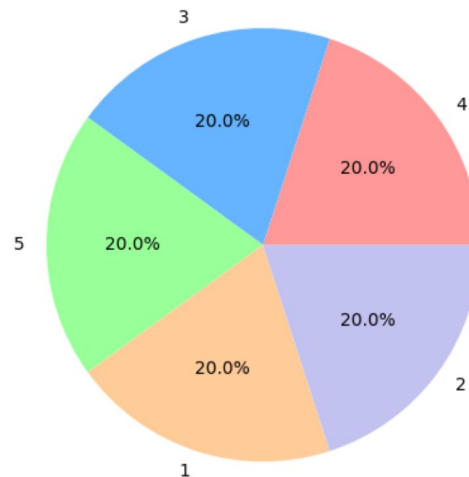
**Review of 5.** Something like this movie won't happen again. This is literally the best superhero movie ever! Thank u MCU

# Preprocesamiento

4. Balanceo del *dataset* previo a entregarle los datos al modelo: misma cantidad de reseñas por cada puntaje



719.514 entradas



280.000 entradas



# División del conjunto de datos

- Se toma el 10% del conjunto para **validación**.
- El objetivo es evaluar a los modelos con datos que **nunca** le fueron **presentados**.





# 04

# Entrenamiento

Técnicas utilizadas y análisis de  
hiperparámetros



# Técnicas

- Congelamiento de capas
- Uso de GPU
- *Fine-tuning* sobre un modelo preentrenado
- Versión *distil* de los modelos
- Tamaño de lote (cantidad de entradas que se entrenan en paralelo)
- Acumulación de gradiente (actualizar varios pasos en lugar de uno por uno)
- Precisión mixta (guardar ciertas variables en la mitad del espacio)

# Hiperparámetros

## Épocas

Nº de veces que se presenta el *dataset* de entrenamiento

## *Learning Rate*

Qué tan grandes serán las actualizaciones de los pesos durante el entrenamiento

## AdamW $\beta_1$ y $\beta_2$

Parámetros del optimizador Adam W

## Batch Size

Cantidad de elementos que forman parte del lote de entrenamiento



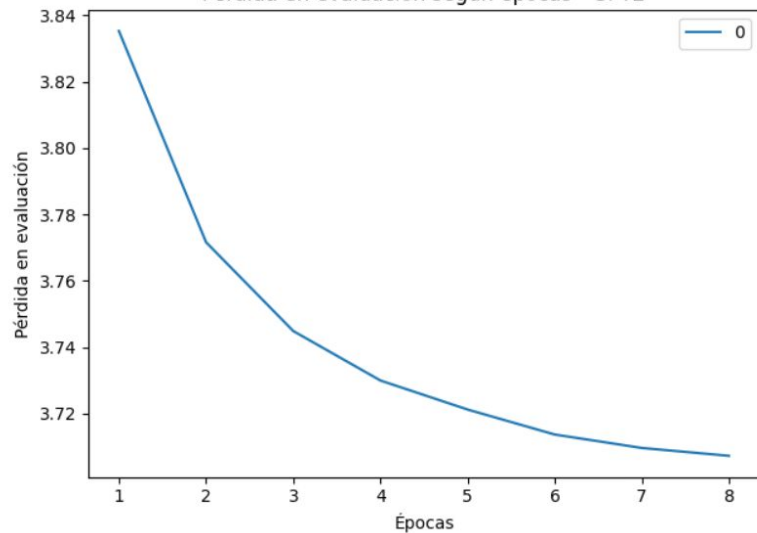
# Consideraciones

- Para el estudio de Hiperparámetros :
  - Tres **muestras disjuntas** de 50.000 entradas cada una
  - Las 3 muestras se utilizan para **cada valor** de un hiperparámetro a analizar
  - **Misma cantidad** de entradas por puntaje
  - Se **promedian** los 3 resultados
  - 20% de datos para **evaluación**
  - Varía únicamente el **hiperparámetro analizado**

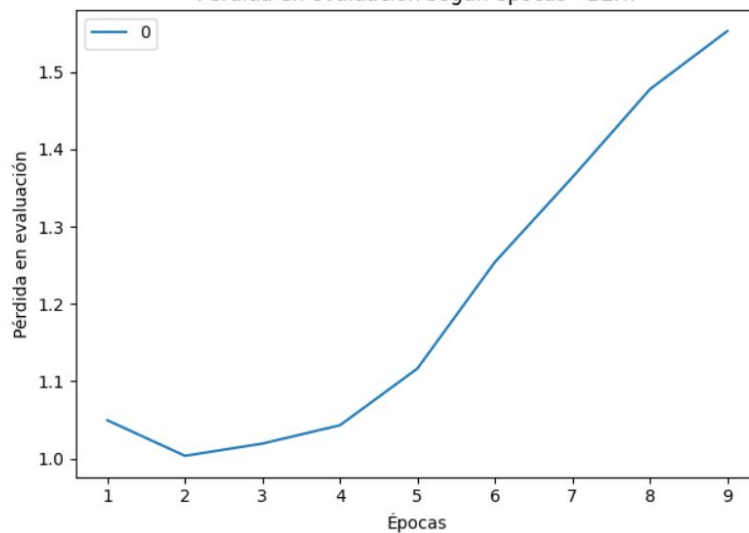


# Épocas

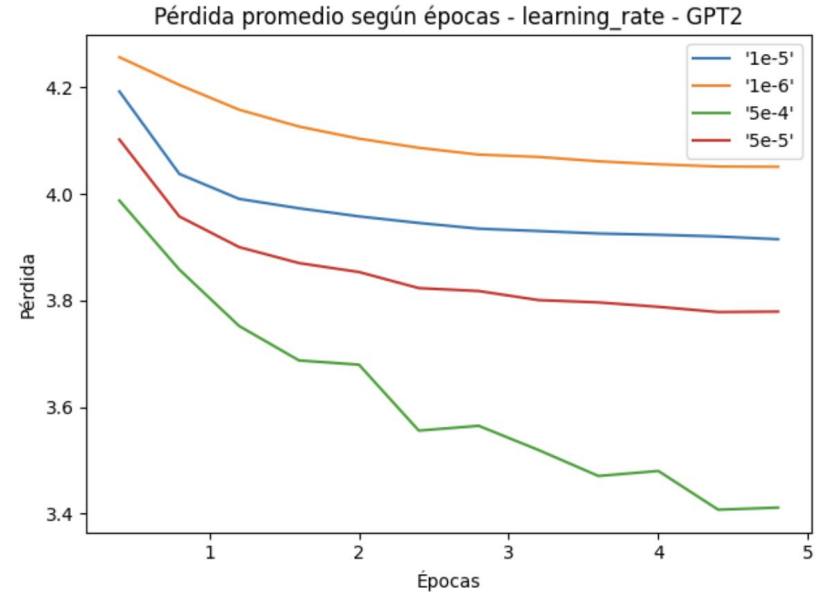
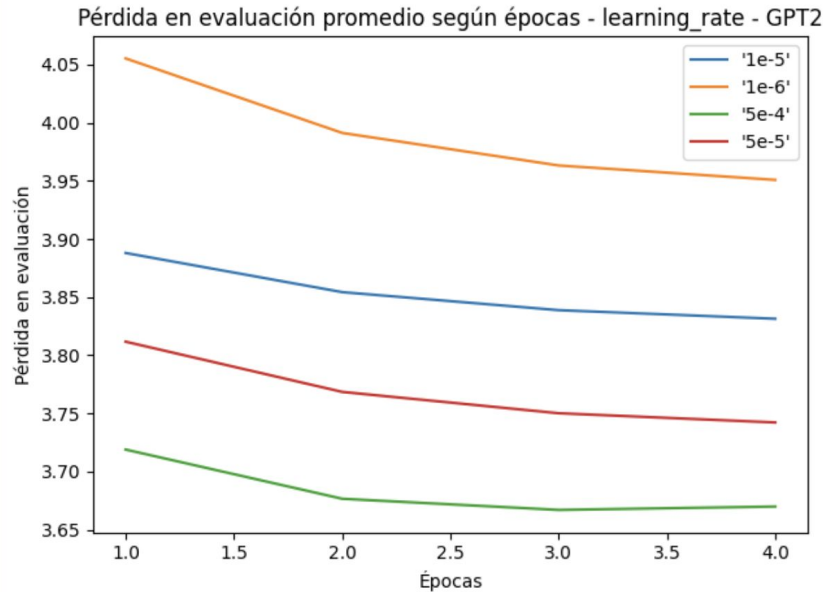
Pérdida en evaluación según épocas - GPT2



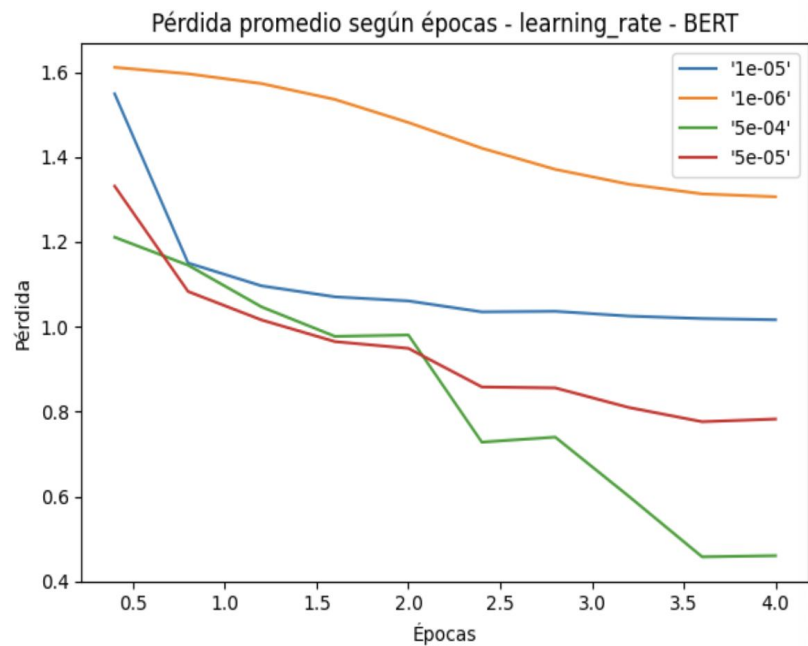
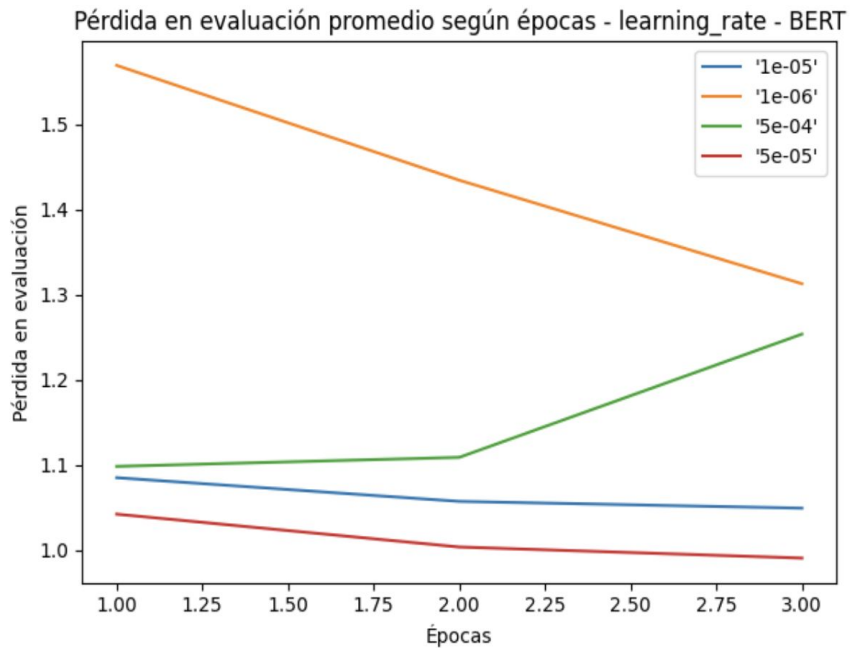
Pérdida en evaluación según épocas - BERT



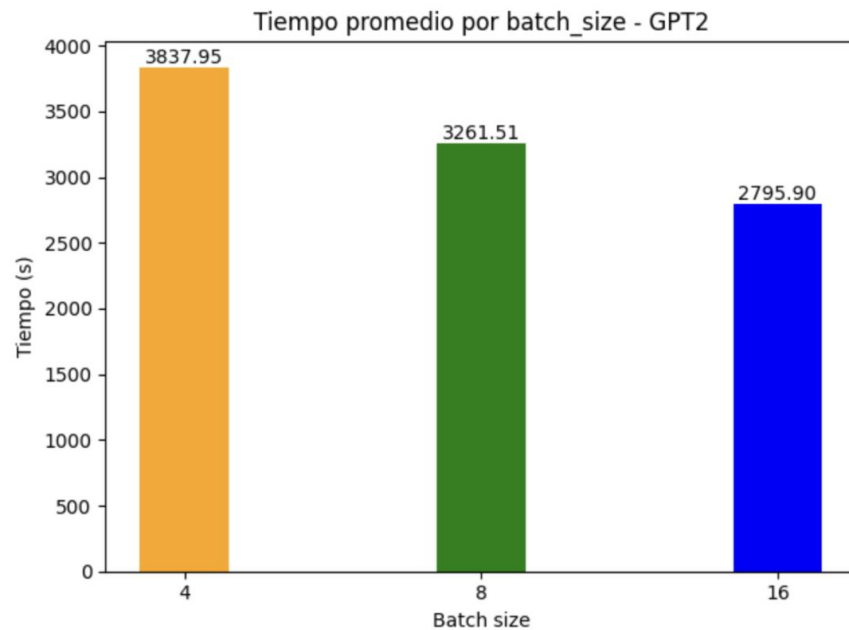
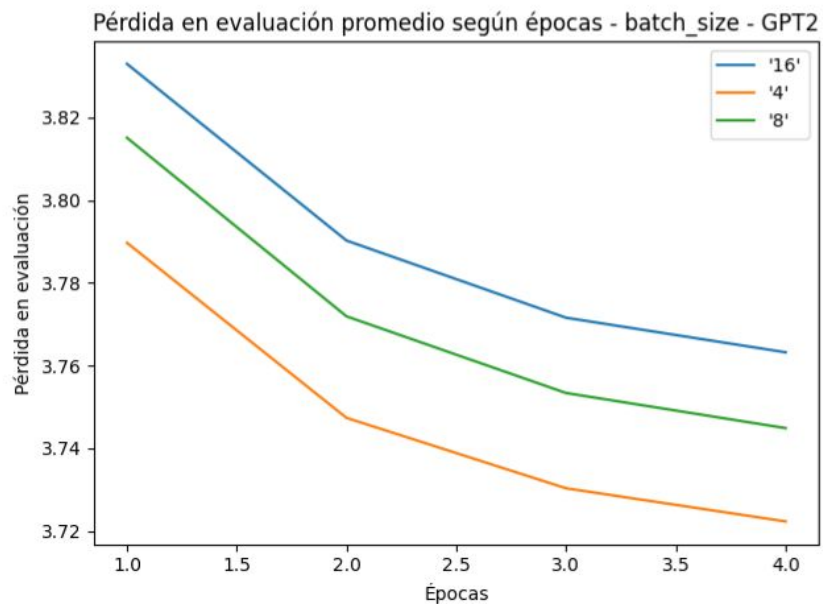
# Learning Rate - Generador



# Learning Rate - Clasificador



# Batchsize - Generador





# 05

## Modelos Finales

Hiperparámetros óptimos

# Hiperparámetros finales

Hiperparámetro	Valor
evaluation_strategy	epoch
num_train_epochs	<b>5 (GPT-2)   3 (BERT)</b>
log_level	error
save_strategy	epoch
fp16	true
per_device_train_batch_size	<b>8</b>
per_device_eval_batch_size	<b>8</b>
gradient_accumulation_steps	<b>4</b>
load_best_model_at_end	true
optim	adamw_torch
learning_rate	<b><math>5 \times 10^{-4}</math> (GPT-2)   <math>5 \times 10^{-5}</math> (BERT)</b>
lr_scheduler_type	linear
weight_decay	0.1
adam_epsilon	$1 \times 10^{-8}$
adam_beta1	<b>0.95</b>
adam_beta2	<b>0.999</b>
disable_tqdm	true
overwrite_output_dir	true
warmup_ratio	0.1
do_eval	true



# Consideraciones: Modelo Final

- Para obtener el modelo **final**:
  - Se entrena **cada red** con 100.000 datos.
  - Se utilizan **conjuntos disjuntos** para entrenar cada red.
  - Se utiliza el 10% del conjunto para **evaluación**.
  - Se obtienen métricas sobre el conjunto de **validación**.



# Métricas

- **Cross Entropy Loss:** función de pérdida de los modelos.
  - Menor valor, las predicciones de la red concuerdan con las reales.
- **Perplexity:** Mide que tanto le sorprende al modelo la existencia de una secuencia.
  - Menor perplexity, menos le sorprende a la red la secuencia



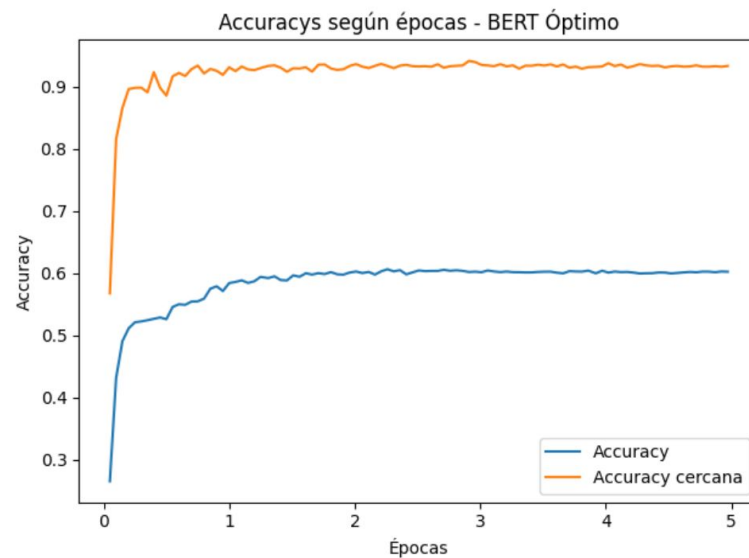
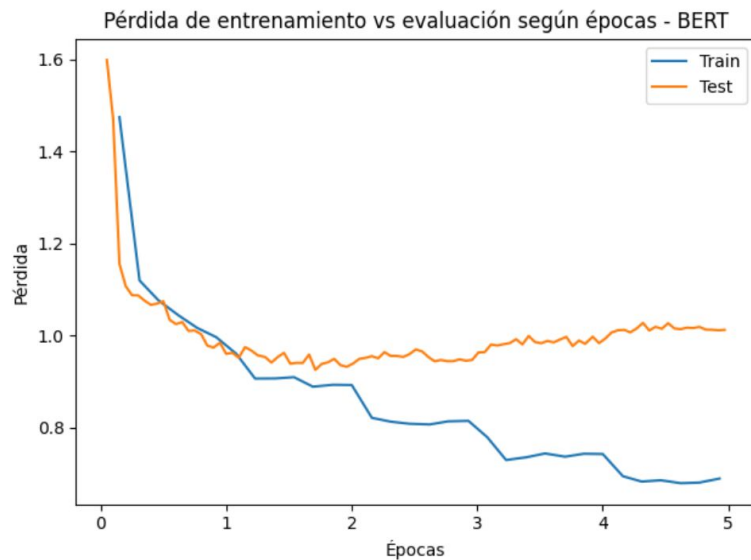


# Métricas cercanas

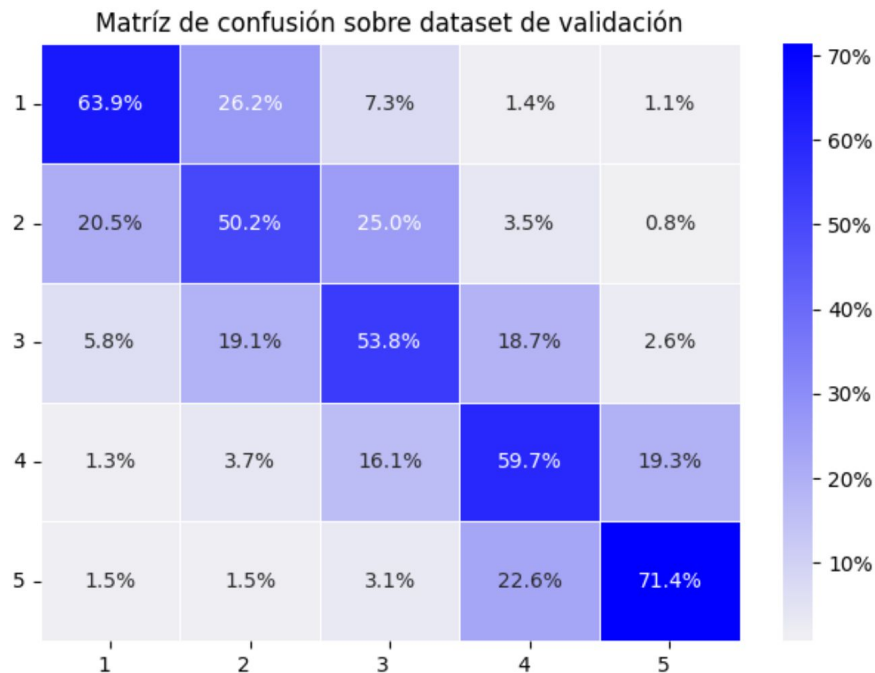
- Se considera **correcta** una predicción del modelo si se encuentra a menos de 1 de **distancia** de la etiqueta real
- R: etiqueta **real**
- P: etiqueta **predicha**

$$P = \begin{cases} R, & \text{si } |P - R| \leq 1 \\ P, & \text{en otro caso} \end{cases}$$

# Validación - Clasificador



# Validación - Clasificador



Métrica	Original	Cercana
Accuracy	60.31 %	93.17 %
Recall	60.31 %	93.17 %
Precision	59.78 %	93.19 %
F1-Score	59.89 %	93.16 %

# Validación - Generador

Sobre el  
conjunto de  
validación

Perplexity	Valor
distilgpt2 <b>Inicial</b>	61.10
distilgpt2 <b>Entrenado</b>	46.28

Sobre 2.500  
reseñas  
generadas

Métrica	Original	Cercana
Accuracy	43.92 %	82.64 %
Recall	43.92 %	82.64 %
Precision	41.97 %	84.58 %
F1-Score	41.39 %	82.28 %



06

# Conclusiones



# Conclusiones

- Queda demostrada la **efectividad** de la técnica implementada en un modelo generador para abordar el problema de la **disponibilidad limitada** de datos.
- Al evaluar el clasificador óptimo obtenido sobre el conjunto de validación, **supera** al clasificador **aleatorio** para las 5 clases y alcanza una tasa de acierto considerable (93.17 %, teniendo en cuenta **errores menores o iguales** a 1).
- Al obtener métricas similares sobre el *dataset* **generado**, se puede concluir que los datos generados por la red son confiables y útiles para el propósito específico del problema abordado.



# Trabajos Futuros

- Realizar entrenamientos con modelos con **mayor cantidad de parámetros**, utilizando técnicas de optimización como PEFT.
- Evaluar las técnicas implementadas en **otro conjunto** de datos (evaluación de desempeño de empleados, noticias, etc.).
- Las matrices de confusión obtenidas sugieren que una división en **menor cantidad de clases** (como positivo y negativo) proporcionaría mejores resultados, dado que los errores se encuentran en las **clases intermedias**.



# Demo

- Se generan 20 reseñas en base a un puntaje deseado
  - Si ninguna fue clasificada de manera correcta, se generan 20 más hasta que haya al menos una predicción correcta.
- Se guardan en archivos diferentes las predicciones correctas de las incorrectas.
- El objetivo es que el clasificador condicione la generación, para obtener datos más confiables.





# DEMO



# Reseñas de puntaje 1

- Bien clasificada
  - Terrible, awful, overrated, overlong, unentertaining drivel, with so many cliches and subplots as well with such an amazing cast of characters that are just terrible to watch. With almost zero story development the movie is in complete limbo.
- Mal clasificada (2)
  - To begin with, this was not as good as I expected it to be. The story felt like a video game or the end of Star Wars. After all, what the last 20 minutes of this episode were? The characters were in a great way, I suppose

## Reseñas de puntaje 2

- Bien clasificada
  - I don't know why they had such high ratings. But I think there is a good reason. The storyline was incredibly slow and boring.
- Mal clasificada (3)
  - This film tries a little while making itself very clever, with some of the best actors in the film, but even though it is not that great on paper. The characters are developed very well in the film.



# Reseñas de puntaje 5

- Bien clasificada
  - This is one of my all time most loved movies and it truly stands out from other Pixar based movies.It's been a long time since I have watched a movie that I consider to be as good as this.
- Mal clasificada (3)
  - I was intrigued by this film.It looked like a true documentary about an era in that era of the big blockbuster movies and it didn't deliver exactly what the real history of war was.



# ¡Muchas gracias!

Matías Lombardi - Patrick Dey

Tutor: Francisco Pérez Sammartino