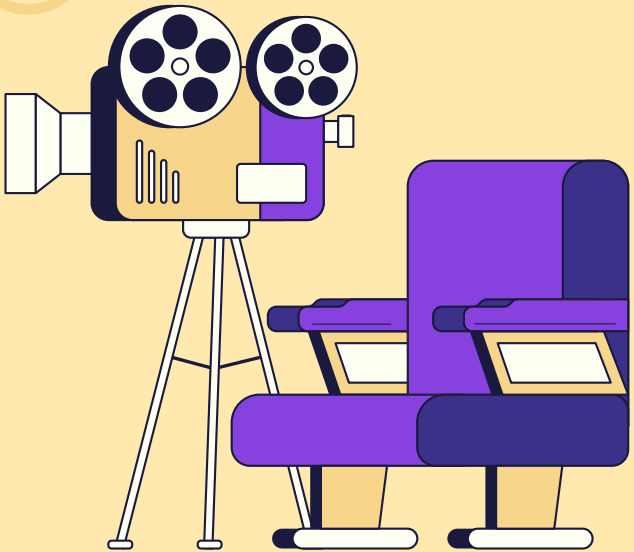
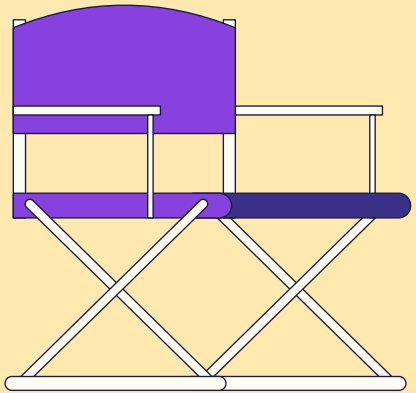


Clasificador de películas

Procesamiento del lenguaje natural

Dey, Patrick
Gordyn Biello, Guido
Rodríguez, Martina
Tarantino Pedrosa, Ana



¿Qué hicimos?

- Procesamiento de un corpus de películas imdb
- Aprendizaje de patrones para generar un modelo clasificador
- Objetivo: recibir una reseña y clasificarla de 1 a 5
- Modelo **distil-bert**
- Comparación con modelo **roberta-base**

Corpus

Características

De terceros

- Películas de IMDb

3 atributos principales

- Nombre de película
- Reseña
- Puntaje

Puntajes

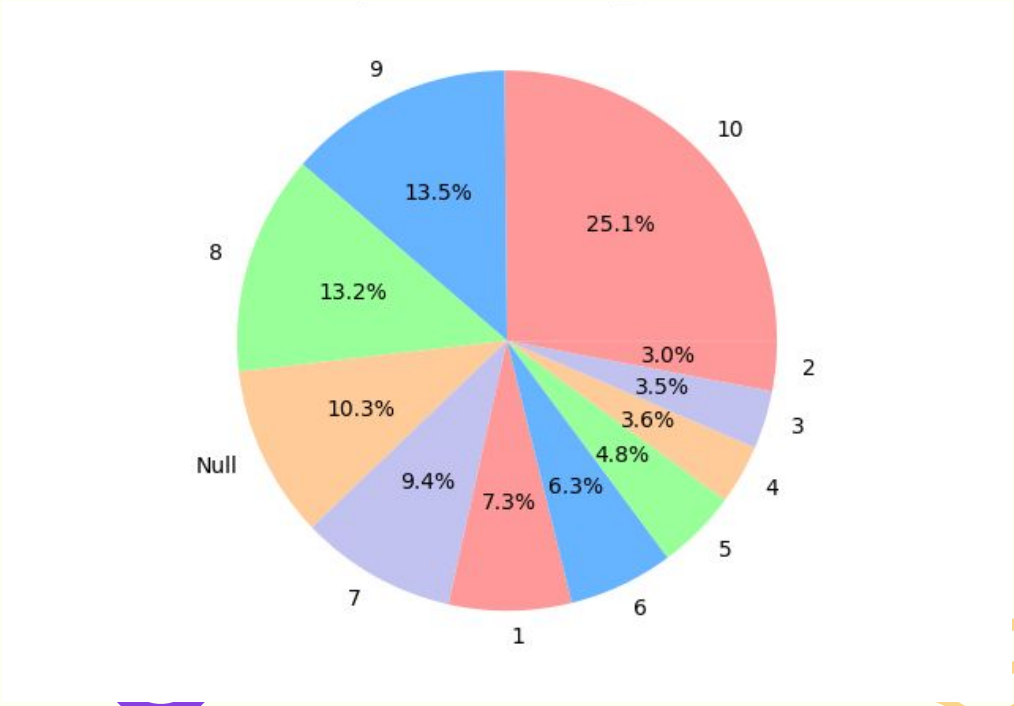
- Reducción de 1-10 a 1-5
- Menos ambigüedad
- 1-10 confusión al clasificar

Preprocesamiento

- Eliminación de **stop-words**
- Trimmeo de espacios, signos de puntuación y exclamación
- Eliminación de tags **HTML** y links
- Reemplazo de **;** a **,**
- Reducción de etiquetas (de 10 a 5 estrellas)
- Se truncan todas las reseña a 512 tokens

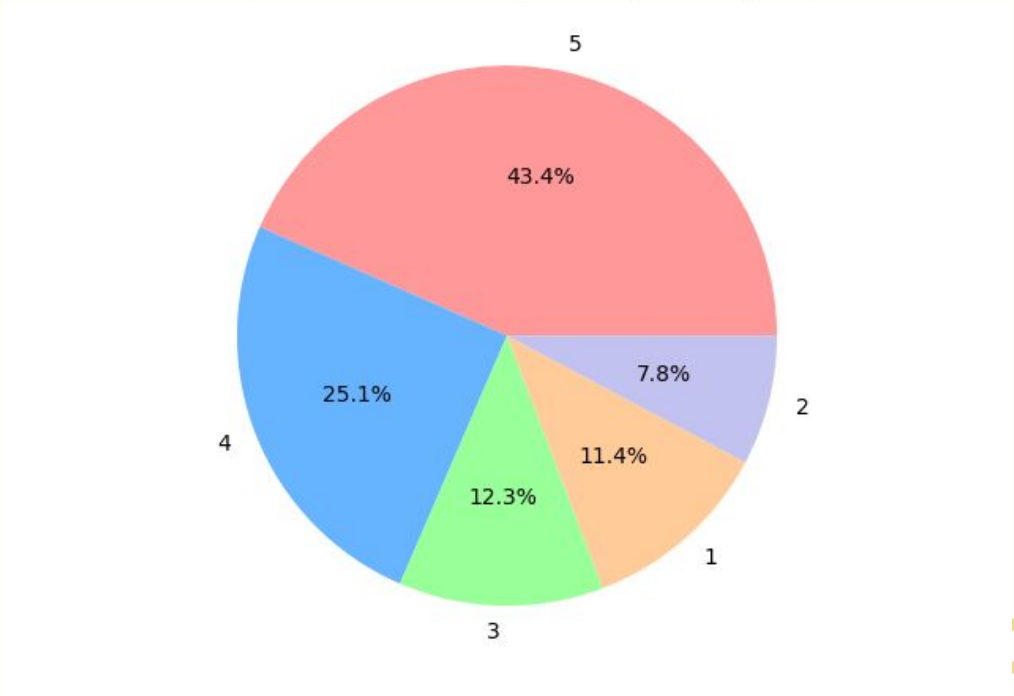
Distribución de las calificaciones

- Antes de la reducción y de remover las *null*



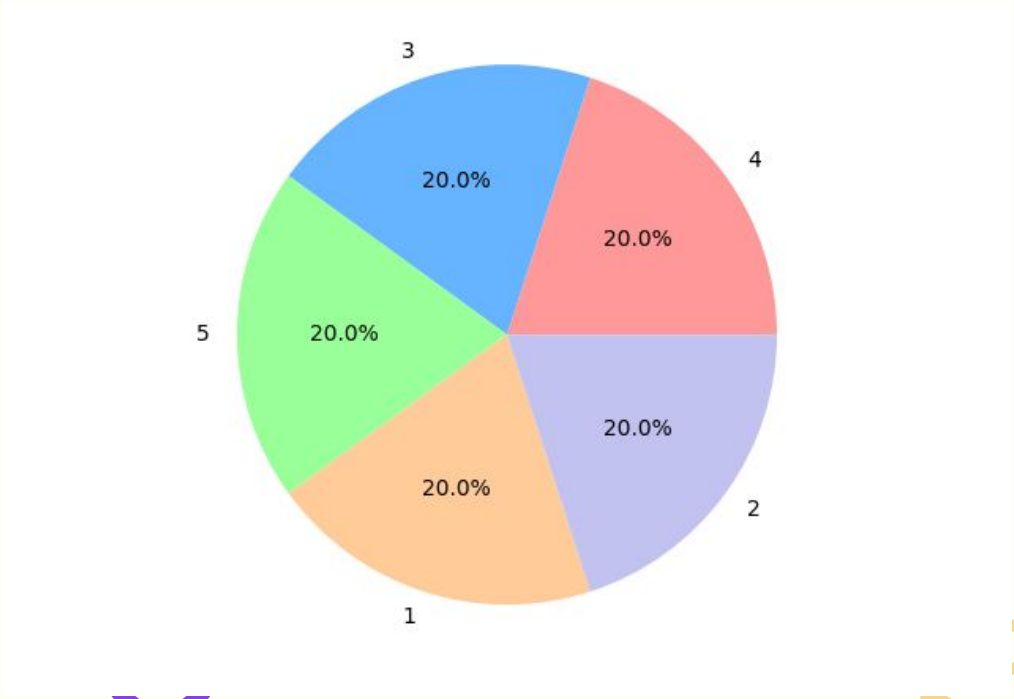
Distribución de las calificaciones

- Luego de la reducción



Distribución de las calificaciones

- Luego del balanceo



53.176

Reseñas por puntaje luego del balanceo

División del dataset

Separación 80-20 para entrenar y luego testear.

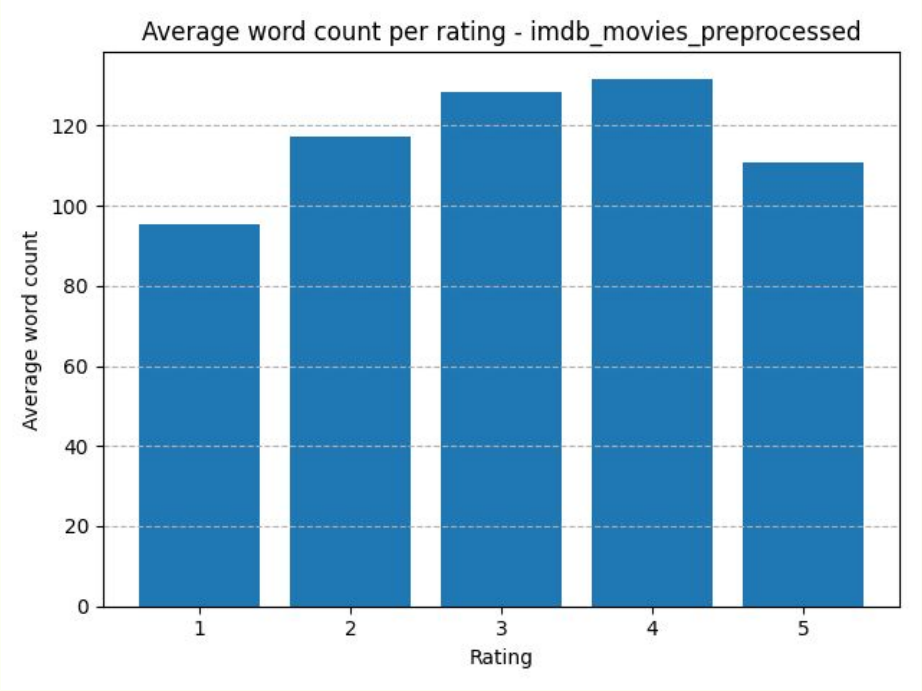
Training Dataset

- 212.700 entradas en total
- 42.540 por puntaje

Test Dataset

- 53.180 entradas en total
- 10.636 por puntaje

Análisis del corpus



Modelos

DistilBERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Utilizado para predicción de palabras enmascaradas
- Fine-tuned para clasificar (utilización del token [CLS])
- Bidireccionalidad

DistilBERT

- Checkpoint: **distil-bert-uncased**
- Truncamos a 512 tokens (máximo)
    - Secuencias de menor tamaño son paddeadas con un token especial
- Asignación de [PAD] token

roBERTa

- Robustly Optimized BERT Approach
- Optimizar y mejorar la arquitectura de BERT
- No se calcula la posición exacta de palabras enmascaradas
- El tamaño del batch puede variar en cada iteración

roBERTa

- Checkpoint: **roberta-base**
- Truncamos a 512 tokens (514 de máximo)
- Utilización del token especial <s> para padding

Entrenamiento

Entorno de trabajo

- Entrenamiento de los modelos
- Utilización de GPU para acelerar procesamiento
- Evitar el uso de los propios recursos

Configuración y uso

- HuggingFace
  - AutoModelForSequenceClassification
  - DataCollatorWithPadding
  - Trainer y Training Arguments
- Argumentos default salvo por
  - Fp16 -> acelerar entrenamiento
  - 5 épocas
  - adamw_torch

Argumentos

```
Code training_args = TrainingArguments(
        evaluation_strategy="epoch",
        num_train_epochs=5,
        log_level="error",
        output_dir="out/bert/imdb80-checkpoints/",
        save_strategy="epoch",
        fp16=True,
        per_device_train_batch_size=16,
        gradient_accumulation_steps=4,
        load_best_model_at_end=True,
        optim="adamw_torch",
        learning_rate=2e-5,
        lr_scheduler_type="linear",
        weight_decay=0.1,
        adam_epsilon=1e-8,
        adam_beta1=0.9,
        adam_beta2=0.999,
        disable_tqdm=True,
        overwrite_output_dir=True,
        warmup_ratio=0.1,
        do_eval=True,
)
```

Resultados

Clasificar una reseña

- Tokenizer el prompt (mismo checkpoint)
- Truncation a 512 tokens
- model(input_ids)
- Softmax sobre los scores devueltos por el modelo
- Devolvemos la etiqueta más probable

$$\sigma : \mathbb{R}^K \to [0,1]^K$$
$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{para } j = 1, \ldots, K.$$

Métricas

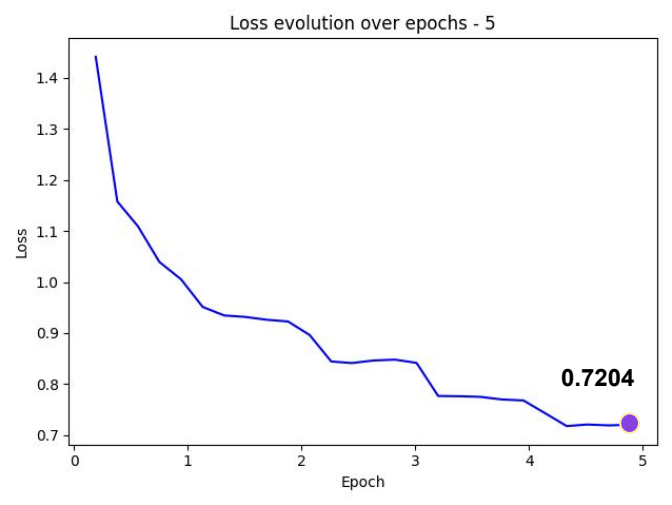
- Se clasifica una reseña y se compara con la etiqueta que ya tenía asignada
- Curva de aprendizaje del modelo (**eval_loss** dada por el modelo)
- Matriz de confusión entre etiquetas asignadas y las generadas
- Evolución de precisión y accuracy época a época

$$\text{Precision} = \frac{TP}{TP + FP}$$

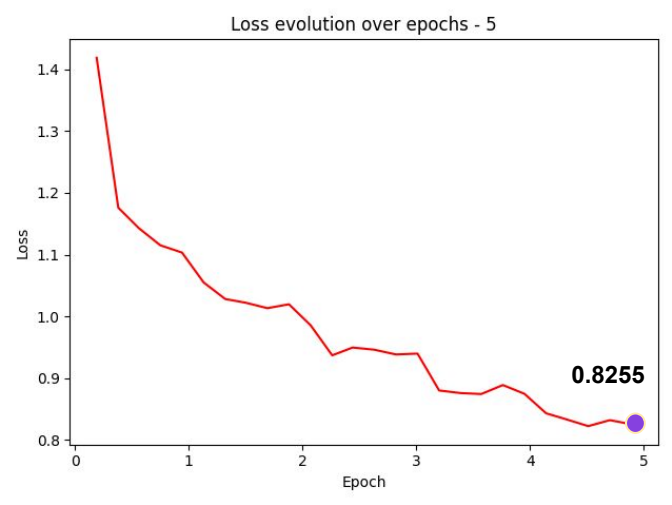
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Curva de aprendizaje

BERT

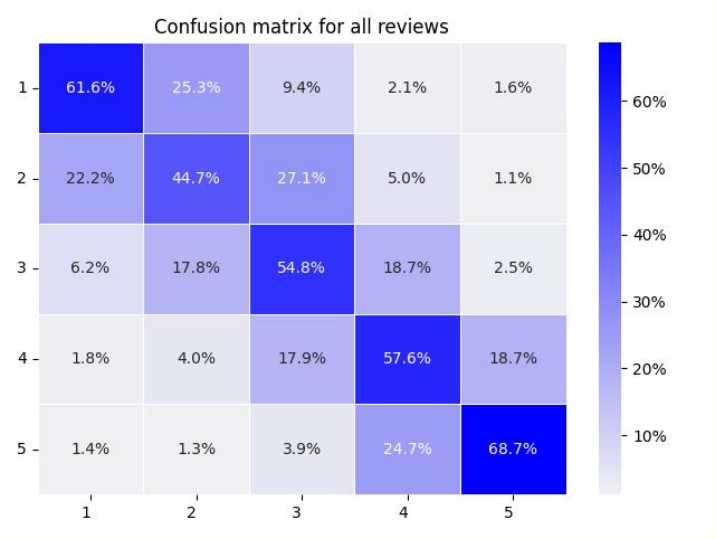


RoBERTa

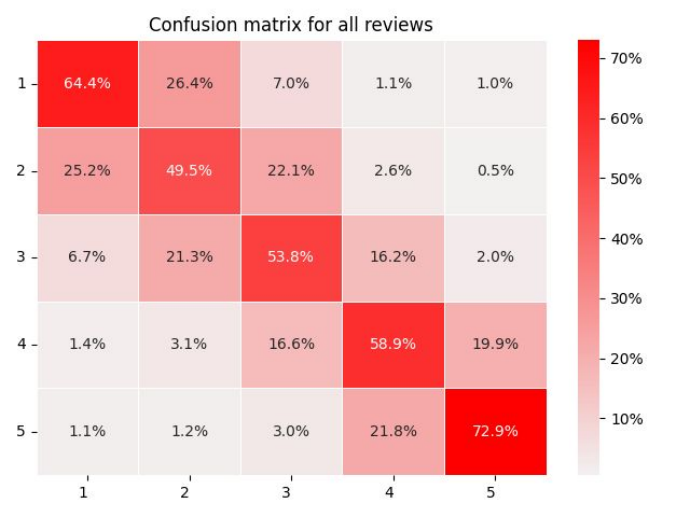


Matriz de confusión

BERT

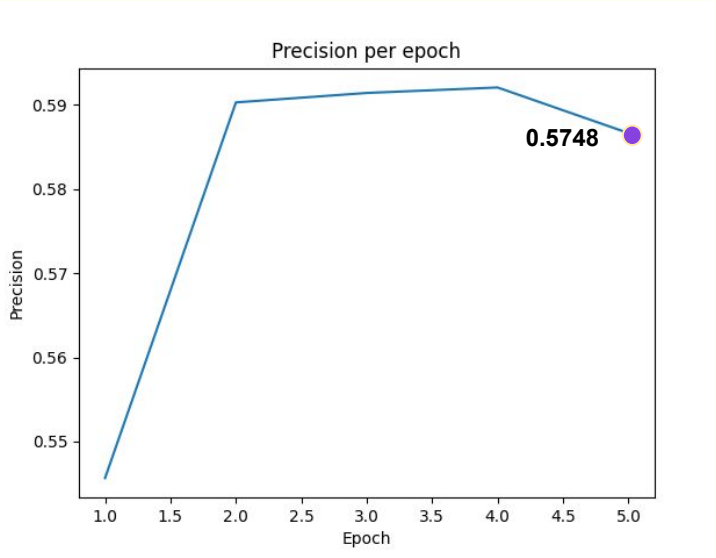


RoBERTa

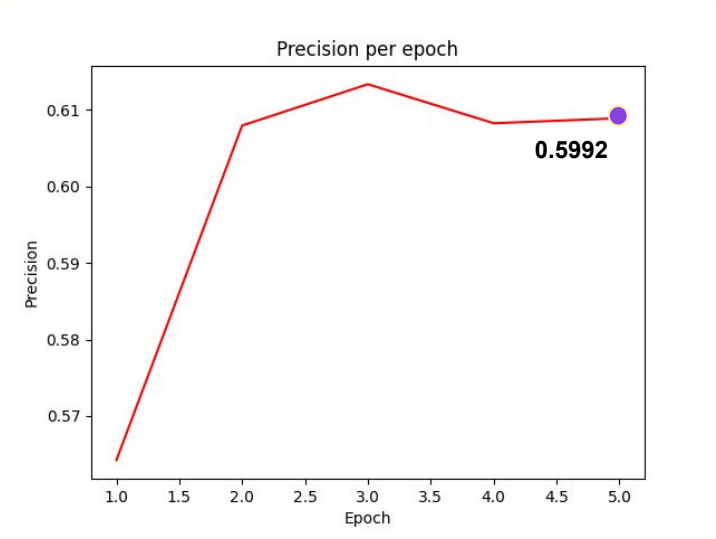


Evolución de la Precisión

BERT

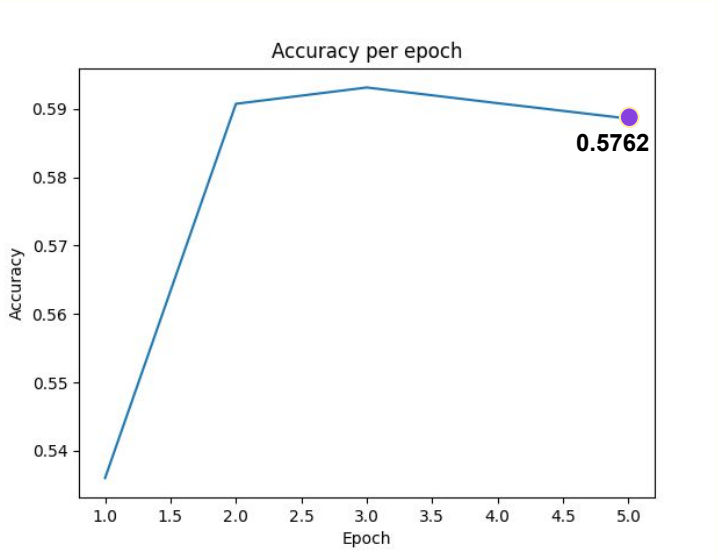


RoBERTa

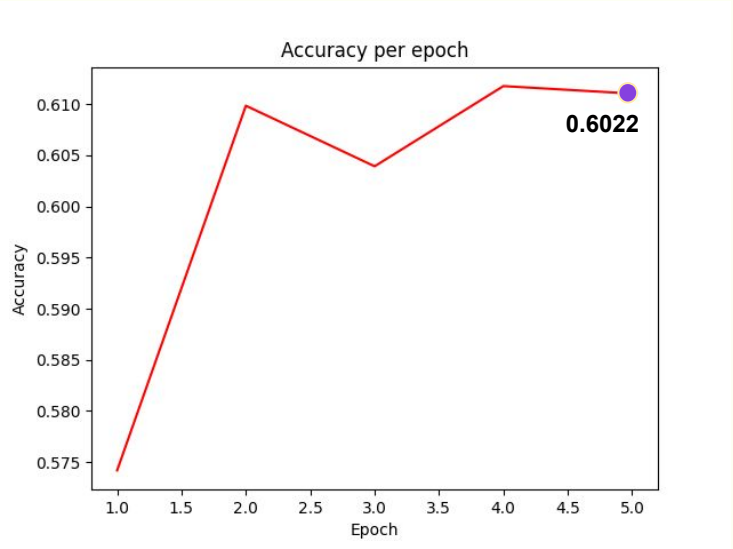


Evolución de la Accuracy

BERT



RoBERTa



Problemáticas

- Falta de K-cross para obtención de métricas debido a la lentitud del entrenamiento
- Comparar métricas con el dataset de entrenamiento
- Definición de la partición de los datos de manera arbitraria
- Dificultad para seleccionar el modelo y comparación

Demo

Algunas reseñas

- When Evil Lurks is a shocking yet highly accomplished horror film, and Rugna's name is one that demands to be up there with Cronenberg, Romero, Carpenter and Argento – this is a true master of the genre.

- In fact, this film had a tendency to overdo it, and often just didn't achieve the intended effect that was expected.I am baffled by people saying this film couldn't even have been cut.While the plot was very intriguing, I was sure it would have easily fallen into the trap of being over, rather than getting rid of it.

- It was an interesting film with good acting by Robert De Niro, but the story was poorly told.This film had none in common with other mafia films.I don't recommend that this film be seen as a cheap watch

¡Muchas gracias!