TP NLP

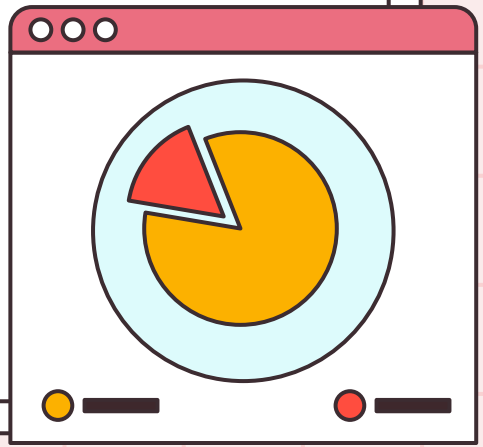
Predicción de géneros de películas

Realizado por:

- Nicolás Casella
- Uriel Arias

1

Objetivos



Objetivos principales

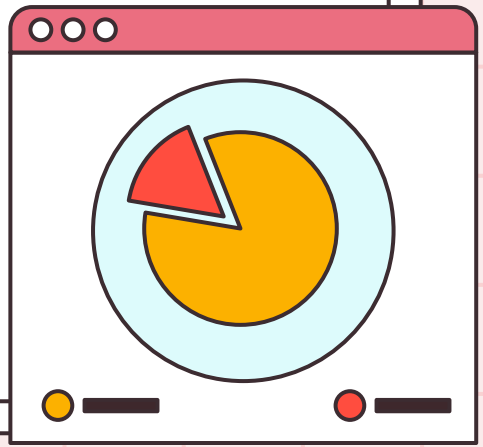
1. Buscar predecir los géneros a los que pertenece una película en base a su sinopsis escrita en inglés.
2. Comparar resultados para clasificación agregando o quitando el título a la sinopsis de la película/documental.
3. Encontrar correlación entre palabras y los géneros de películas.

Posibles usos

- Catalogar películas/documentales por estrenarse.
- Recomendaciones de películas/documentales.

2

Metodología



Preprocesamiento general

Dataset original

- **letterboxd-all-movie-data**: Dataset con información de películas en Letterboxd, disponibilizado a través de Hugging Face.

Transformaciones

- Se descartan las features no utilizadas como url, cast, directors, etc.
- Se eliminan las películas que no tienen género o sinopsis y las filas repetidas.
- Se filtran los 18 géneros más populares.

Dataset filtrado

Campos utilizados

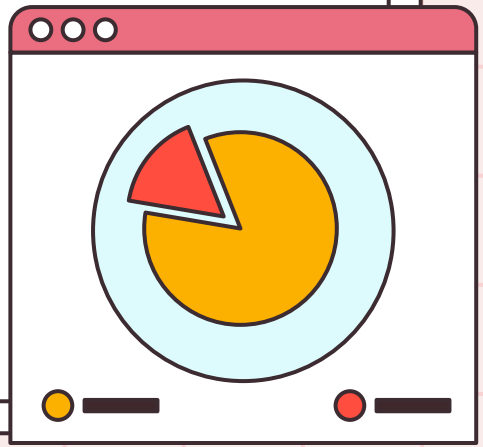
- **synopsis** (str)
- **genres** (str[])
- **title** (str) → luego preappendeado a la sinopsis

Géneros filtrados

```
"Drama", "Documentary", "Comedy", "Animation",
"Horror", "Romance", "Thriller", "Music",
"Action", "Crime", "Family", "Adventure",
"Fantasy", "Science Fiction", "Mistery", "History",
"War", "Western"
```

2

Experimentos y resultados



Experimento: Clasificación multiclase

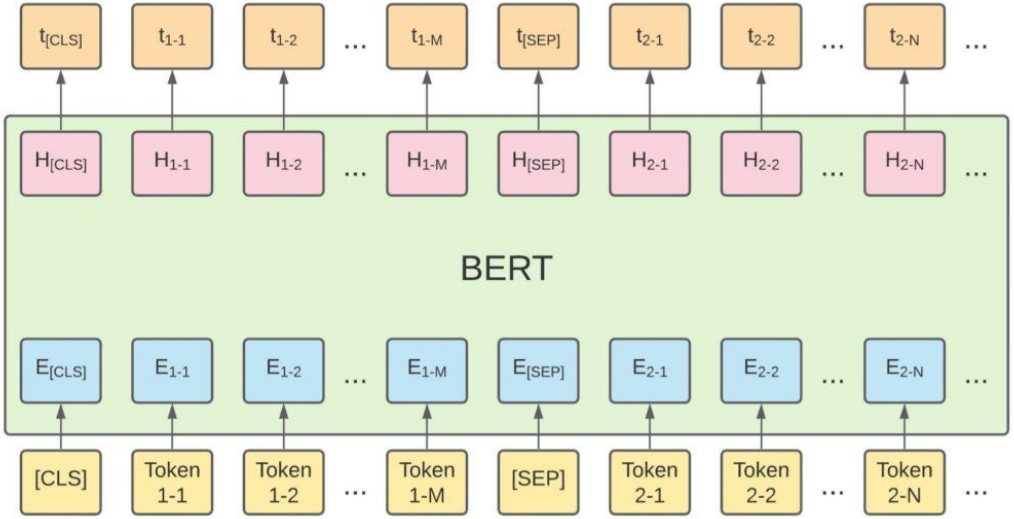
Dataset utilizado

- **multi-genre:** Dataset filtrado con sinopsis de películas que pertenecen de 1-3 géneros. Se muestran la misma cantidad de ejemplos por género (1 género) y de películas agrupadas por número de géneros (5100 filas).

Herramienta utilizada

- **Modelo de base**: BERT-base-cased disponibilizado por Hugging Face.
- **Técnica utilizada**: Finetuning con dataset muti-genre.
- **Data Split**: 80-10-10 (training, validation, test).

BERT-base-cased



- Modelo encoder only.
- ~ 100M de parámetros.
- Sin capas agregadas

Finetuning

Hiperparámetros

- **epochs**: 5
- **learning_rate**: 1e-4
- **per_device_train_batch_size**: 8
- **per_device_eval_batch_size**: 8
- **weight_decay**: 0.01
- **metric**: "macro_f1"

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

$$MacroF1 = \frac{1}{N} \sum_{i=1}^{N} F1_i.$$

Salida del modelo

Mapeo para la elección de géneros

- Se pasaron los logits a una función logística para recuperar las probabilidades y luego se filtraron los valores > 0,5.

```
[[0. 0. 0. ... 0. 0. 1.]
 [0. 0. 0. ... 0. 1. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [1. 0. 0. ... 1. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

Vectores esparsos con
len = cantidad de géneros

Resultados: clasificación solo según sinopsis

Mejor entrenamiento

Epoch	Training loss	Validation Loss	Macro F1-Score
1	No log	0.2988	0.4736
2	0.0423	0.2898	0.5105
3	0.0323	0.2945	0.5277
4	0.0323	0.2973	0.5405
5	0.0115	0.3009	0.5367

Resultados: clasificación solo según sinopsis

Ejemplo exitoso:

Cantidad de géneros	Sinopsis	Géneros predichos	Géneros reales
1	This short film offers a glimpse into the life of Louis Hippolyte Lafontaine, the Chief Justice who died prematurely but left French Canada a legacy of political freedom. Shot entirely in Montreal, the film begins on the day of his death, and flashes back to tense moments throughout his life. French with English subtitles...	['History']	['History']

Resultados: clasificación solo según sinopsis

Ejemplo exitoso:

Cantidad de géneros	Sinopsis	Géneros predichos	Géneros reales
2	Zombie Prom is a 1950s horror comic book brought to life as a musical comedy film. It is a campy, rollicking, romp through America's "Atomic Age" and the "Golden Age" of horror comic books. (Fragmento)	[‘Comedy’, ‘Horror’]	[‘Comedy’, ‘Horror’]

Resultados: clasificación solo según sinopsis

Ejemplo “semiexitoso”:

Cantidad de géneros	Sinopsis	Géneros predichos	Géneros reales
3	When a young man new in town, the impressionable Chris Donds (Esteban Powell), is drawn into the super wealthy scene of Los Angeles after dark [...] His new lifestyle of carousing and daredevil brushes with the law leads up to an entanglement with the Russian mafia. (Fragmento)	['Crime', 'Drama']	['Comedy', 'Crime', 'Drama']

Resultados: clasificación solo según sinopsis

Ejemplo de fracaso:

Cantidad de géneros	Sinopsis	Géneros predichos	Géneros reales
2	Zara loses her child due to miscarriage. Bilal and Zara decide to adopt a child but they don't know what is going to happen to them… (Fragmento)	['Drama']	['Horror', 'Thriller']

Resultados: clasificación según sinopsis y título

Epoch	Training loss	Validation Loss	Macro F1-Score
1	0.30880	0.268551	0.259568
2	0.243100	0.245227	0.437146
3	0.18970	0.240637	0.523205
4	0.140200	0.243817	0.539558
5	0.098400	0.248602	0.568782

Resultados: clasificación según sinopsis y título

Ejemplo exitoso:

Cantidad de géneros	Sinopsis	Géneros predichos	Géneros reales
3	The Wizard of Oz: Young Dorothy finds herself in a magical world where she makes friends with a lion, a scarecrow and a tin man as they make their way along the yellow brick road to talk with the Wizard and ask for the things they miss most in their lives… (Fragmento)	[‘Adventure’ ,'Family', 'Fantasy']	['Adventure', 'Family', 'Fantasy']

Resultados: clasificación según sinopsis y título

Ejemplo de fracaso :

Cantidad de géneros	Sinopsis	Géneros predichos	Géneros reales
2	Doggy Doggo Dinner Diner: Zed Ramos presents a short film telling an interview between a dog and the oldest living Filipino… (Fragmento)	['Comedy']	['Drama', 'Fantasy']

Experimento: NER sobre sinopsis

Dataset utilizado

- **one-genre:** Dataset filtrado solo con sinopsis de películas que pertenecen a un solo género (19500 filas).

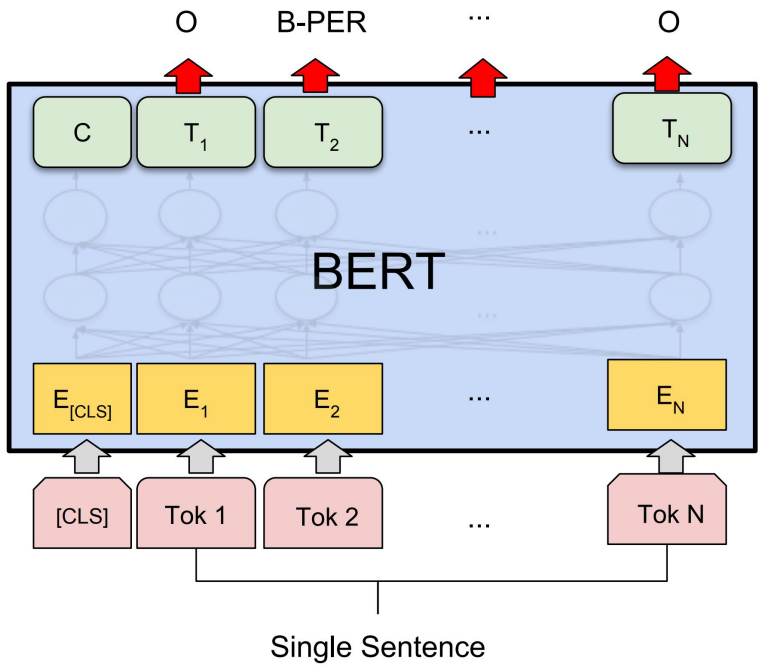
Herramienta utilizada

- **Modelo de prueba**: BERT-base-NER disponibilizado por Hugging Face.
- **Técnica utilizada**: Inferencia (sin finetuning previo)

NER

Entidad	Significado
O	Entidad no nombrada
B-MISC	Comienzo de una entidad miscelánea que le sigue a otra entidad miscelanea
I-MISC	Entidad miscelánea
B-PER	Comienzo del nombre de una persona justo después del nombre de otra persona
I-PER	Nombre de persona
B-ORG	Comienzo de una organización justo después de otra organización
I-ORG	Organización
B-LOC	Comienzo de una ubicación justo después de otra ubicación
I-LOC	ubicación

BERT-base-NER



- Modelo BERT cased finetuneado con el dataset de NER en inglés.
- ~100M de parámetros

Postprocesamiento NER

- **Simplificación de entidades**: Se agrupan las entidades que marcan el inicio de una entidad con el final de la misma y se eliminan las entidades desconocidas (O).
  - Entidades postprocesadas:
    - LOC
    - PER
    - ORG
    - MISC
- **Filtrar según score**: se filtran entidades con score < 0,8.

Resultados: Top entidades LOC destacadas

Western

Texas (83), **California** (59), **Arizona** (58), **Mexico** (53), **West** (43)

Documentary

America (111), **Europe** (83), **Russia** (56), **United States** (55)

Science Fiction

Earth (10)

Animation

Earth (68), **Japan** (26), **China** (17), **New York** (13)

Resultados: Top entidades PER destacadas

Animation

Woody (64)

Western

Tom (112), **Jim** (99), **Bill** (96), **Steve** (82)

Thriller

Zach (5), **Dee** (4), **Wolf** (4)

Resultados: Top entidades ORG destacadas

Documentary

Edison (11), **UN** (9), **WWE** (9), **Google** (9)

Music

Grateful Dead (103), **MTV** (41), **BBC** (28)

Comedy

Comedy Central (8), **HBO** (6)

Animation

NFB (National Film Board, 16), **Academy Film Archive** (13), **UCLA Animation Workshop** (7)

Western

Army (26), **Texas Rangers** (23), **Pony Express** (14), **Wells Fargo** (14)

Resultados: Top entidades MISC destacadas

Drama

French (11), **American** (9), **Japanese** (8)

Horror

Zombie (3)

Animation

Japanese (59), **American** (32), **French** (23)

Western

Western (178), **Indian** (170), **Mexican** (152), **Civil War** (67)

Experimento: correlación de géneros con adjetivos de la sinopsis

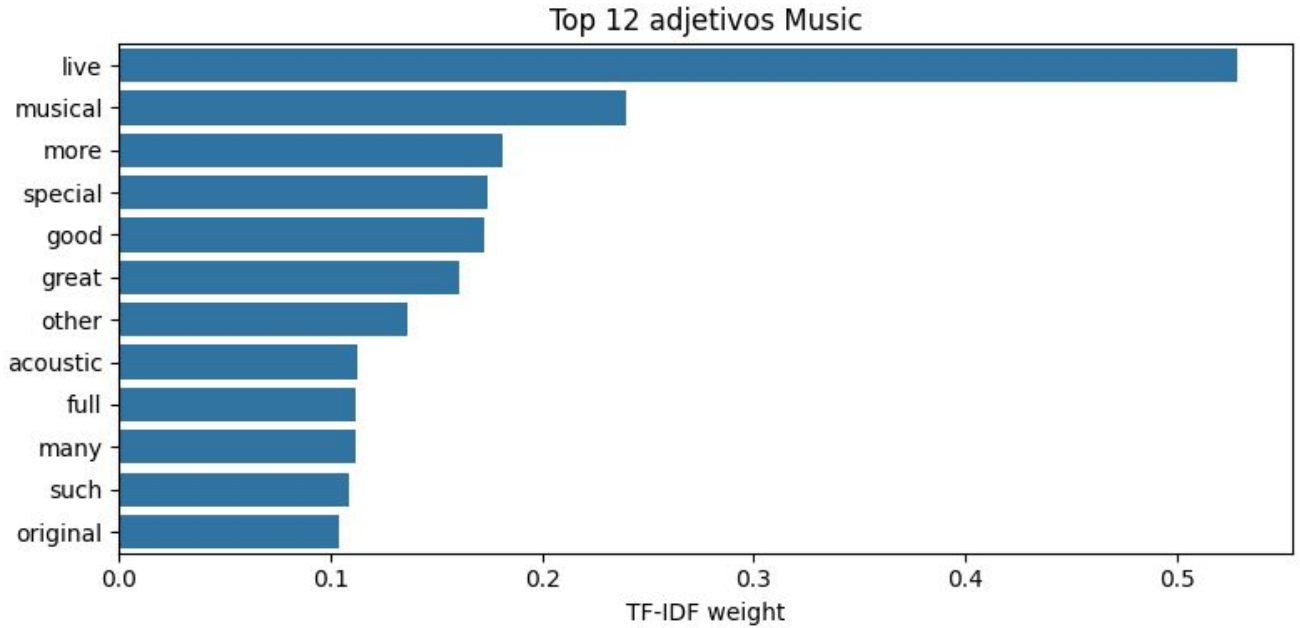
Dataset utilizado

- **one-genre:** Dataset sanitizado filtrado con sinopsis de películas que pertenecen a un solo género.

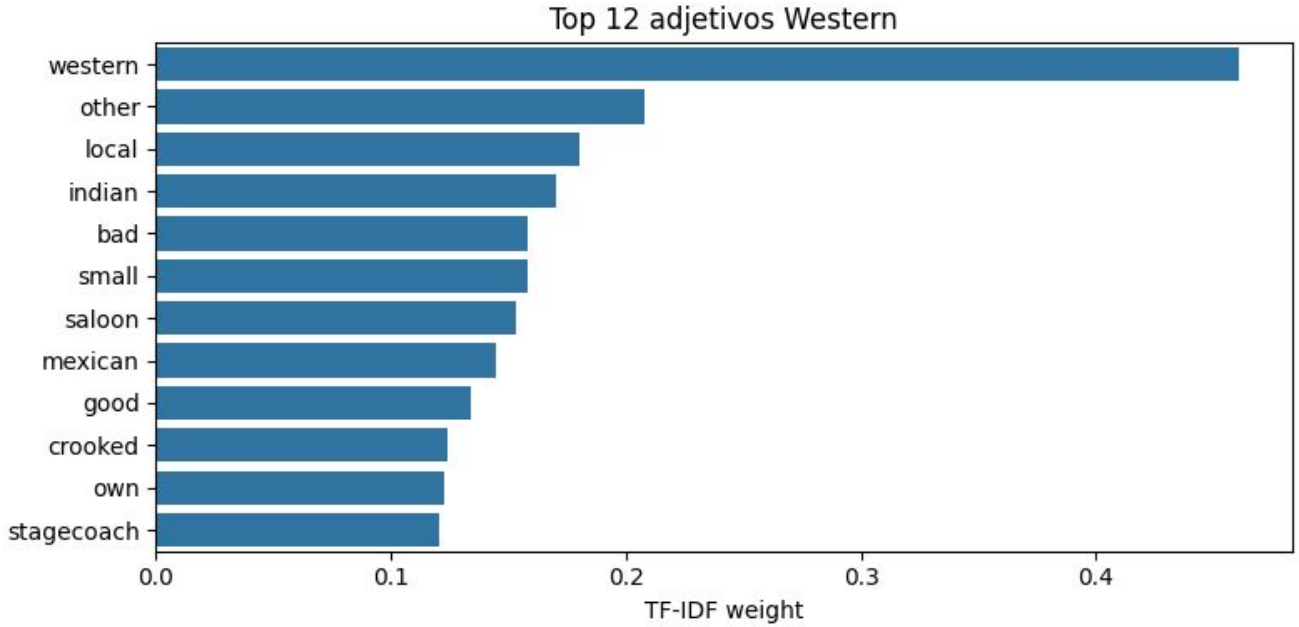
Herramienta utilizada

- **Librería SpaCy**: Pipeline de filtrado de adjetivos.
- **Técnica utilizada**: TF-IDF.

Resultados: correlación de géneros con adjetivos de la sinopsis



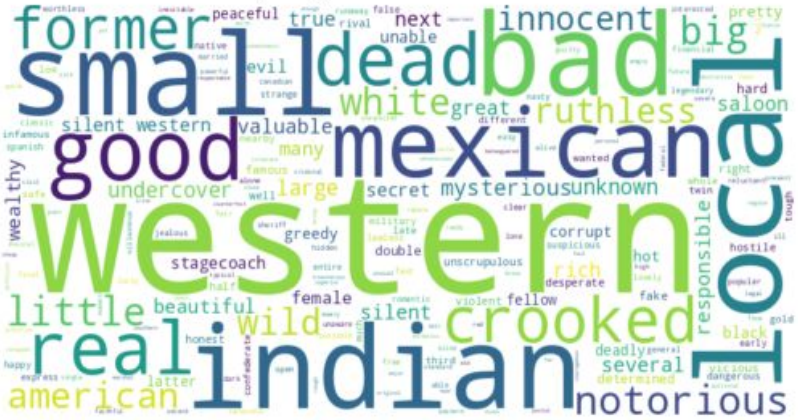
Resultados: correlación de géneros con adjetivos de la sinopsis



Resultados: correlación de géneros con adjetivos de la sinopsis

Top adjetivos en Science Fiction

Top adjetivos en Western



Resultados: correlación de géneros con adjetivos de la sinopsis

Top adjetivos en Horror

Top adjetivos en Romance

Experimento: correlación de performance con longitud de sinopsis

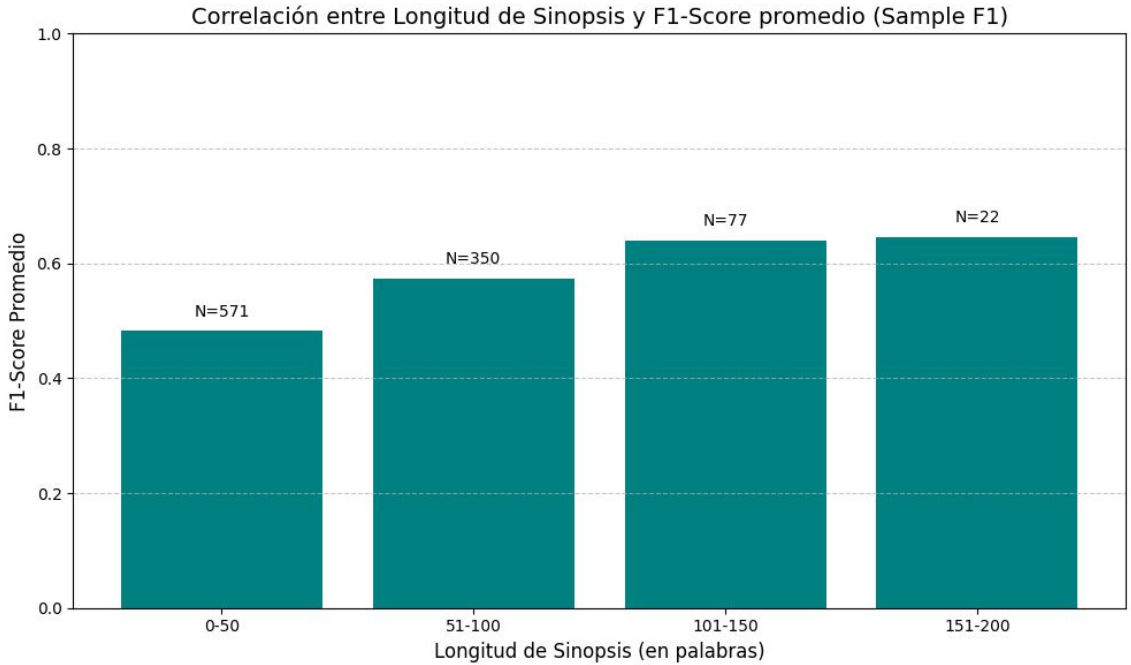
Dataset utilizado

- **multi-genre:** Dataset filtrado con sinopsis de películas que pertenecen de 1-3 géneros. Se muestran la misma cantidad de ejemplos por género (1 género) y de películas agrupadas por número de géneros (5100 filas).

Herramienta utilizada

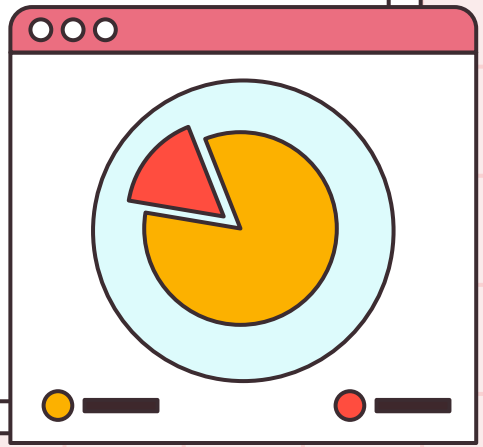
- **Modelo utilizado**: modelo de clasificación finetuneado propio
- **Técnica utilizada**: Inferencia

Resultados: correlación de performance con longitud de la sinopsis



4

Conclusiones

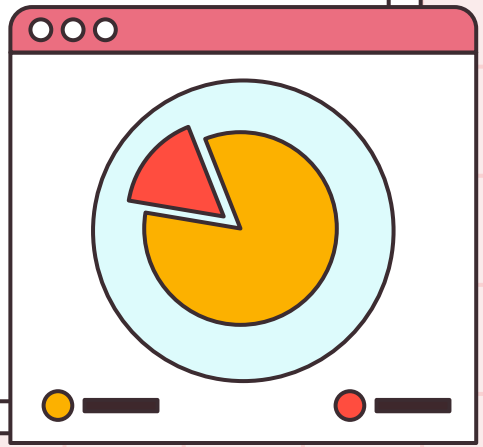


Conclusiones

- Se obtuvo una mejor métrica para la clasificación incluyendo el título a la sinopsis, sin embargo, no fue significativo el cambio.
- Se pudo encontrar una correlación entre ciertas palabras en géneros.
- A mayor cantidad de palabras en la sinopsis, se ve una mejora en la clasificación en función a la métrica f1-score.
- Adicionalmente, se encuentran adjetivos que tienen gran peso en todos los géneros.
- Se pudieron identificar ciertas entidades características para cada género que podrían ser de ayuda para identificar patrones en las sinopsis.

5

Limitaciones y posibles mejoras



Consideraciones finales

Limitaciones

- **Calidad del dataset**: tener mayor variedad de películas que tengan más de un género. Además se requiere revisión humana
- **Capacidad de cómputo:** poder usar modelos con mayor número de parámetros para finetunear.

Mejoras posibles

- Expandir el análisis a otros idiomas.
- Hacer un entrenamiento más exhaustivo explorando otras opciones que optimicen el proceso.

Gracias!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**