



PROYECTO NLP - Libros



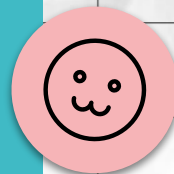
A. C. Luis
alcaluis@alumni.uv.e
s

C. B. Rebeca
recombar@alumni.uv.e
s

M. M. Marta
memuiz@alumni.uv.e
s

V. Alejandra
ave6@uv.es

Grupo 1





Índice

01

Introducción

Web Scraping. Preprocesado. EDA.

02

Modelos

Propios. Pre-entrenados. Comparativa.

03

Conclusiones y trabajo a futuro



1. Introducción

Web Scraping. Preprocesado. EDA.

1. Introducción

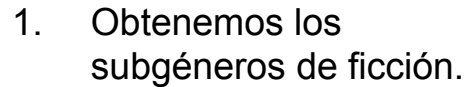
El objetivo de nuestro proyecto es



Sinopsis



**Identificar el
género**



2. Por cada género hacemos una petición a la API de OpenLibrary

3. La api nos devuelve un .json con información de los libros.

1. Introducción - Preprocesado

Hemos realizado la limpieza a distintos niveles:

1. Texto limpio
2. Texto preprocesado
3. Texto preprocesado + lematizado

```
Sinopsis original:
A very real little girl named Alice follows a remarkable rabbit down a rabbit hole and steps through a looking-
--back cover

Contains:

- [Alice's Adventures in Wonderland](https://openlibrary.org/works/OL8193508W)
- [Through the Looking Glass, and What Alice Found There][2]

[2]: https://openlibrary.org/works/OL15298516W

-----
Sinopsis procesada (no lema):
very real little girl named alice follows remarkable rabbit down rabbit hole steps through looking glass come f
-----
Sinopsis procesada (lema):
very real little girl name alice follow remarkable rabbit down rabbit hole step through look glass come face fa
-----
Sinopsis procesada (limpia):
A very real little girl named Alice follows a remarkable rabbit down a rabbit hole and steps through a looking
back cover
```

Librería: Spacy

Modelo:
en_core_web_lg

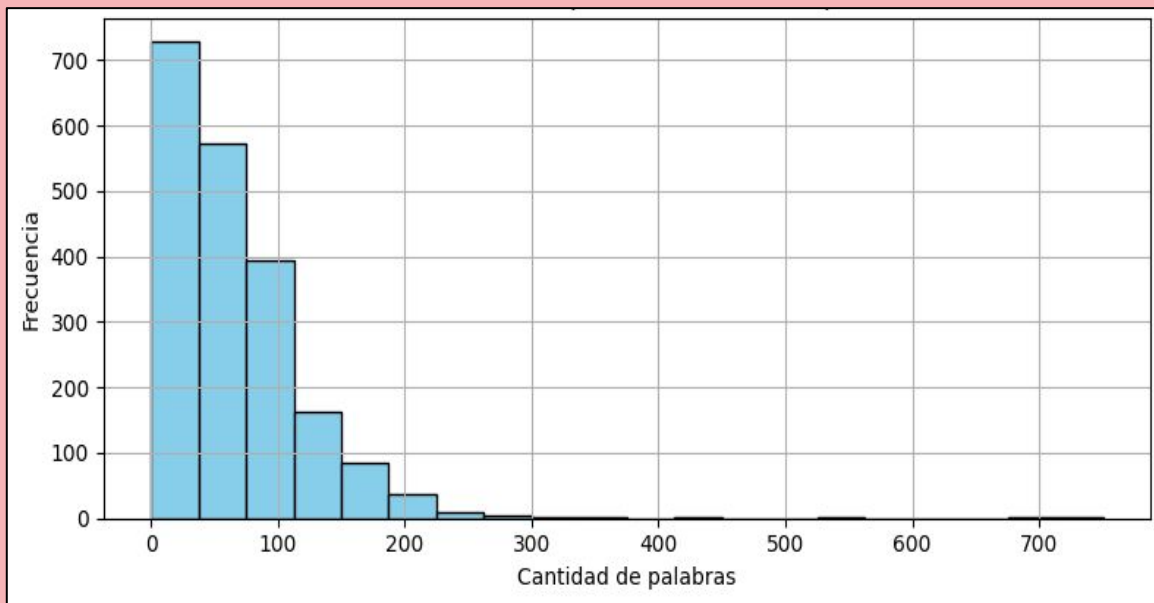
Pasos considerados:

- Eliminación de contenido irrelevante
- Expandir contracciones
- Eliminar puntuación y *stopwords*.
- Conversión a minúsculas

Extracción características:

- BoW, TF-IDF, Embeddings

Distribución de Palabras



- El **90%** de los textos tiene menos de 133 palabras
- Eliminamos libros con menos de 14 palabras que corresponde al **percentil 10**.

1. Introducción - Análisis EDA

Objetivos:

- Comparar el *enfoque propuesto* y el *enfoque original*.
- Evaluar distribución de palabras y de géneros en el dataset.
- Identificar características útiles para clasificación.

Enfoque propuesto



Problema multi etiqueta con géneros agrupados y filtrados.

Eliminación de “géneros” como lo son historias cortas o humor.

Agrupación de géneros con temática cercana como “Magia” y “Fantasía”.

Total: 4 géneros

Enfoque original

Problema de clasificación multi etiqueta donde se clasifican 14 géneros distintos.

Total: 14 géneros

Cantidad libros

Libros
leídos

2000

con sinopsis
largas

1800

con géneros
relevantes

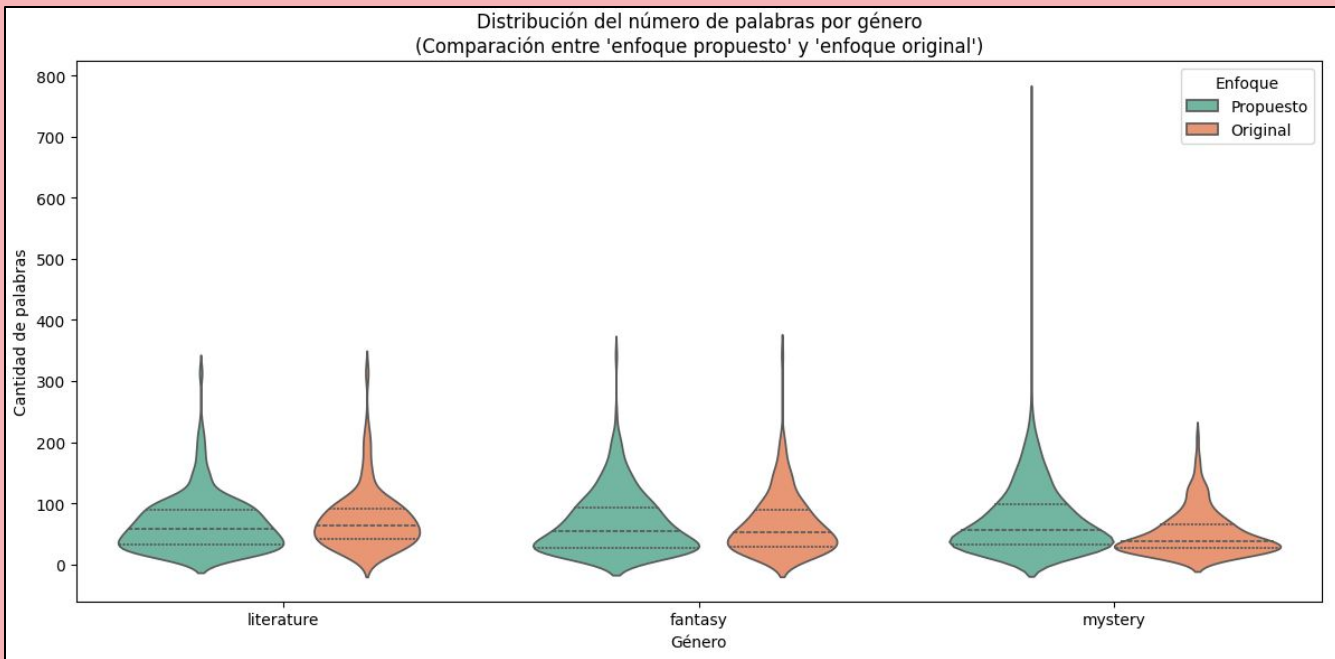
1200

10

[illegible][illegible]

Distribución de Palabras por género

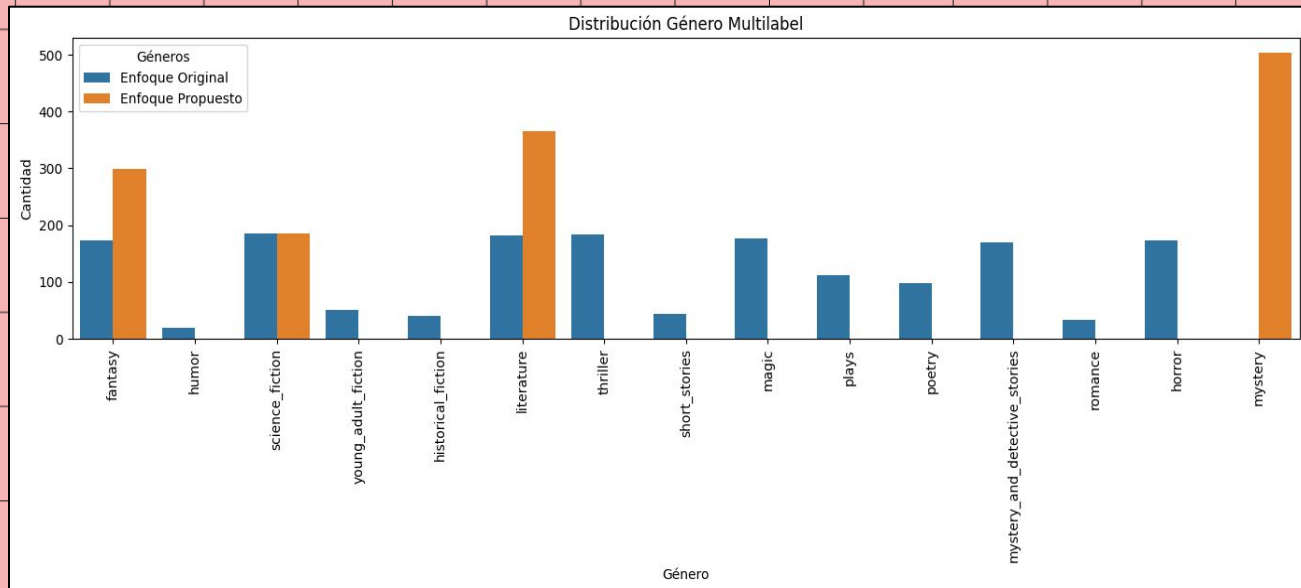
11



- Aumenta la densidad en el nuevo enfoque.
- Las formas de distribución se mantienen similares.
- No permite distinguir géneros entre sí.

Distribución género multilabel

12



- Mystery y Literature agrupan más géneros y concentran más asignaciones.
- Se observa un desbalance en las nuevas categorías.

El gráfico muestra frecuencia de etiquetas, no cantidad de libros únicos.

2. Modelos

Propios. Pre-entrenados. Comparativa.

Modelo propio

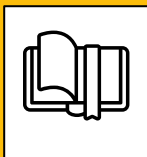
Extracción de características:



Bag Of Words con y sin lematizado



TF-IDF con y sin lematizado



Embeddings (BERT y Sentence embeddings)

Clasificadores:

Regresión logística

XGBoost

SGD Classifier

Estrategias:

MultiOutput

OneVsRest

Umbral dinámico

Mejores resultados para BoW

15

Feature Extraction	Strategy	Classifier	F1 Score (weighted)	Accuracy
BOW_lemma	MultiOutput	LogisticRegression	0.744	0.617
BOW_lemma	OneVsRest	LogisticRegression	0.744	0.617
BOW_no_lemma	MultiOutput	LogisticRegression	0.722	0.528
BOW_no_lemma	OneVsRest	LogisticRegression	0.722	0.528
BOW_lemma	OneVsRest	XGBoost	0.709	0.540
BOW_lemma	MultiOutput	SGDClassifier	0.709	0.565

Mejores resultados para TF-IDF

16

Feature Extraction	Strategy	Classifier	F1 Score (weighted)	Accuracy
TFIDF_lemma	OneVsRest	LogisticRegression	0.768	0.657
TFIDF_lemma	MultiOutput	LogisticRegression	0.768	0.657
TFIDF_no_lemma	MultiOutput	LogisticRegression	0.763	0.629
TFIDF_no_lemma	OneVsRest	LogisticRegression	0.763	0.629
TFIDF_lemma	OneVsRest	SGDClassifier	0.761	0.629
TFIDF_lemma	MultiOutput	SGDClassifier	0.760	0.625

Mejores resultados para embeddings

17

Feature Extraction	Strategy	Classifier	F1 Score	Accuracy
Embeddings_sentence_transformers	OneVsRest	LogisticRegression	0.777	0.653
Embeddings_sentence_transformers	MultiOutput	LogisticRegression	0.777	0.653
Embeddings_sentence_transformers	MultiOutput	XGBoost	0.739	0.605
Embeddings_sentence_transformers	OneVsRest	XGBoost	0.739	0.605
Embeddings_sentence_transformers	OneVsRest	SGDClassifier	0.739	0.593
Embeddings_BERT	OneVsRest	LogisticRegression	0.734	0.577

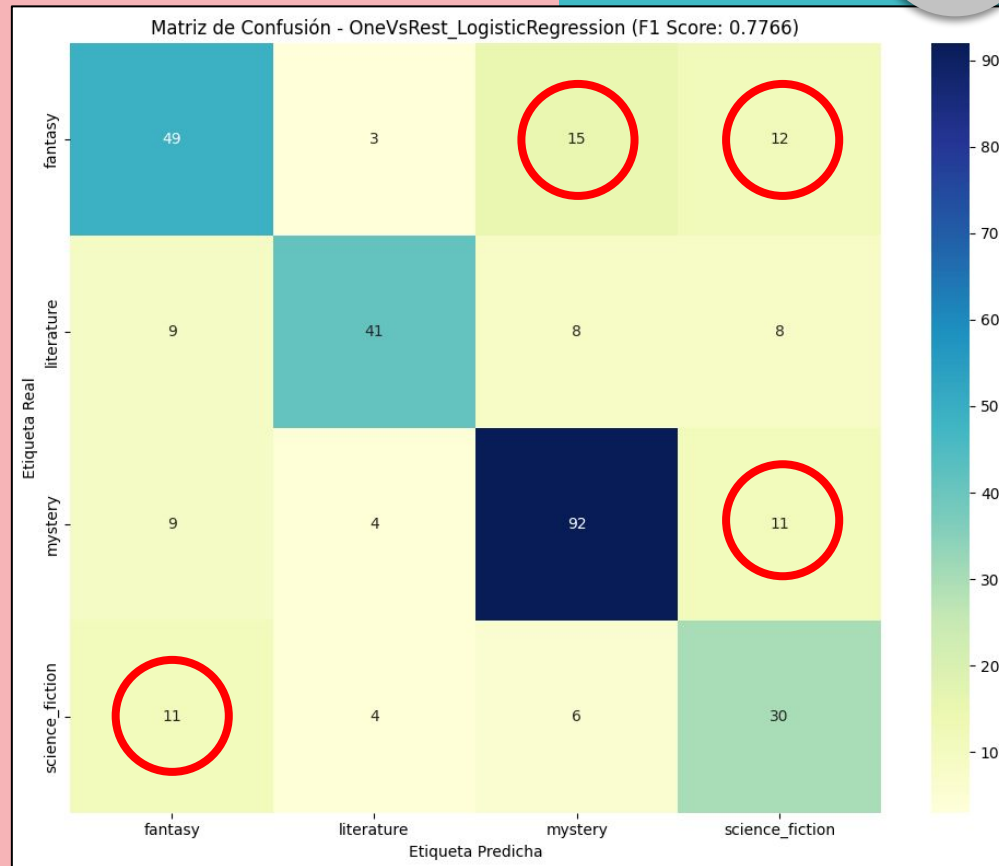
El mejor modelo (propio)

18

	Clasificador	f1 score
Sentence Embeddings	Regresión logística	0.78
TF-IDF con lematizado	Regresión logística	0.77

Entrenando y evaluando: OneVsRest_LogisticRegression
F1 Score (weighted): 0.7766
Accuracy: 0.6532
Classification Report:

	precision	recall	f1-score	support
fantasy	0.75	0.80	0.78	61
literature	0.85	0.64	0.73	64
mystery	0.81	0.88	0.84	105
science_fiction	0.59	0.79	0.67	38
micro avg	0.76	0.79	0.78	268
macro avg	0.75	0.78	0.76	268
weighted avg	0.78	0.79	0.78	268
samples avg	0.76	0.80	0.76	268



Discusión de errores

- Mayor confusión entre los tres géneros más parecidos: fantasía, misterio y ciencia ficción.
- Sobreestima ciencia ficción (posible consecuencia de clases desbalanceadas)
- Asigna misterio en lugar de fantasía (posible por la propia naturaleza de una sinopsis)

Discusión de errores

"For Harry Potter, it's the start of another far from ordinary year at Hogwarts when the Knight Bus crashes through the darkness and comes to an abrupt halt in front of him. It turns out that Sirius Black, mass murderer and follower of Lord Voldemort, has escaped."

Harry Potter y el prisionero de Azkaban.

"In his sixth year, the names Black, Malfoy, Lestrangle and Snape will haunt Harry with shades of trust and treachery as he discovers the secret behind the mysterious Half-Blood Prince."

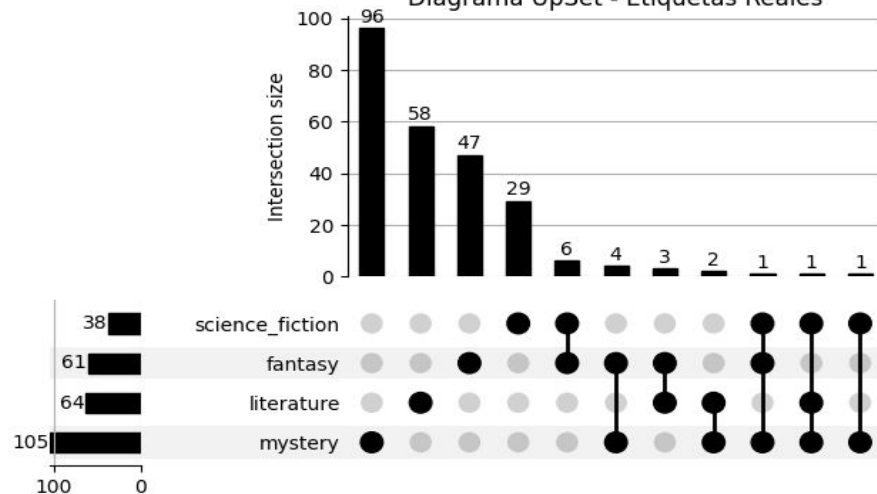
Harry Potter y el príncipe mestizo.

Etiquetas reales:
fantasía, literatura



Etiqueta predicha:
misterio (misterio,
thriller y horror)

Diagrama UpSet - Etiquetas Reales



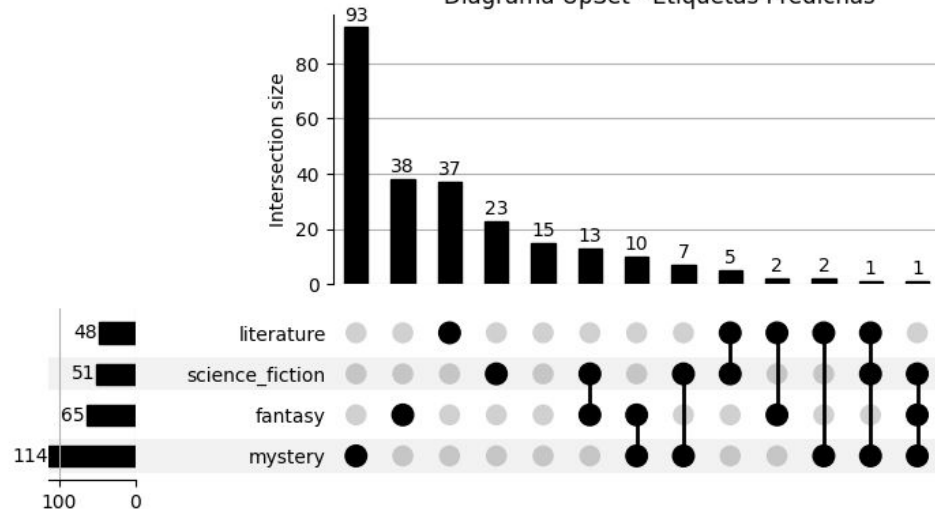
- Modelo tiende a asignar más etiquetas.
- Hay 15 obras de test que no ha clasificado en ninguna.
- Menor número de libros clasificados con literatura como única etiqueta.

HAMLET

Literatura (teatro)


Literatura, fantasía

Diagrama UpSet - Etiquetas Predichas



Modelo pre-entrenado

22



DistilBERT community



[Activity Feed](#) [Follow](#) 261

AI & ML interests

This organization is maintained by the transformers team at Hugging Face and contains the historical (pre-"Hub") DistilBERT checkpoints.

Team members


2





Models


9


Sort: Recently updated


**distilbert/distilbert-base-multilingual-cased**
Fill-Mask • Updated May 6, 2024 • ↓ 10.7M • ♥ 190


**distilbert/distilbert-base-uncased-distilled-squad**
Question Answering • Updated May 6, 2024 • ↓ 202k • ⚡ ♥ 115


**distilbert/distilbert-base-cased** ~66 millones
Fill-Mask • Updated May 6, 2024 • ↓ 217k • ⚡ ♥ 46

**distilbert/distilbert-base** ~82 millones
Fill-Mask • Updated Feb 19, 2024 • ↓ 950k • ⚡ ♥ 155

**distilbert/distilbert-base-uncased-finetuned-sst-2-...**
Text Classification • Updated Dec 19, 2023 • ↓ 4.47M • ⚡ ♥ 765

**distilbert/distilbert-base-german-cased**
Fill-Mask • Updated May 6, 2024 • ↓ 41.3k • ♥ 21

**distilbert/distilbert-base-cased-distilled-squad**
Question Answering • Updated May 6, 2024 • ↓ 224k • ⚡ ♥ 244

**distilbert/distilbert-base-uncased**
Fill-Mask • Updated May 6, 2024 • ↓ 14.4M • ⚡ ♥ 692

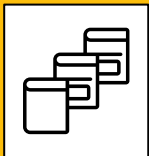
**distilbert/distilgpt2**
Text Generation • Updated Feb 19, 2024 • ↓ 3.32M • ♥ 531

distilbert-base-cased

Versión destilada de BERT



Distingue mayúsculas y minúsculas



BookCorpus y English Wikipedia



Predicción de palabras enmascaradas

Entrenamiento:

Multi Etiqueta

Umbral dinámico

Texto limpio

AutoModel

Modificamos el clasificador

Fine-tuning

Fine-tuning

Fine-tuning basado en características

Congelar embeddings y transformer
Re-entrenar el clasificador

Fine-tuning parcial

Congelar embeddings y 4 capas del transformer
Re-entrenar 2 capas y el clasificador

Fine-tuning total

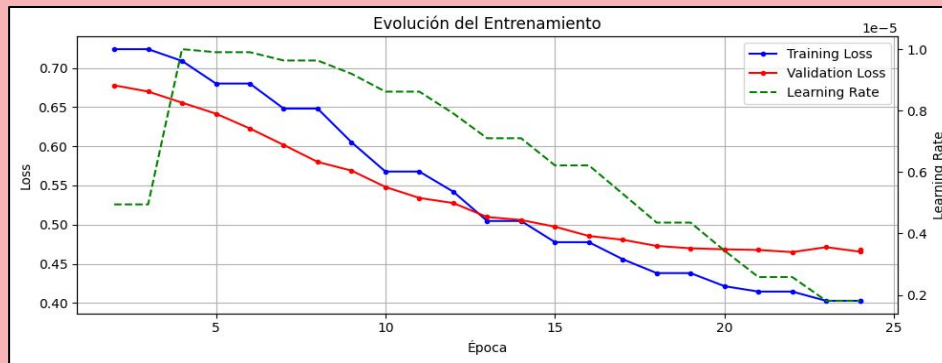
Re-entrenar el modelo completo

Resultados modelo pre-entrenado

25

Fine-tuning	Parámetros	Épocas	F1 Score	Accuracy
Basado en características	691.332	20 (máxima)	0.797	0.710
Parcial	14,867,076	15	0.824	0.754
Total	65,882,244	24	0.816	0.710

Sobreajuste





3. Conclusiones y Trabajo a futuro

Sumario

28

Objetivo

Clasificar a partir de sinopsis

Datos

- Obtención
- Preprocesado
- Análisis

EC

- Enfoque propuesto y original
- BoW, TF-IDF, Embeddings

M. Propio

- Lematizado y no lematizado
- BoW, TF-IDF, Embeddings
- Análisis de la clasificación

M. Pre.

- Comparación niveles de fine-tuning

Conclusiones y Dificultades

29

- Sinopsis
 - No es suficiente para resolver el problema. La calidad es baja. Además de ser pocas palabras.
 - Fácil obtención.
- Etiqueta (Géneros)
 - Gran cantidad de géneros pobremente definidos y entremezclados.
- Comparativa modelos con decentes resultados en ambos enfoques ligeramente superiores en pre-entrenado.
 - M. Propio tiene un dataset pequeño en comparativa con el corpus empleado en el M. pre-entrenado.

M. Propio				M. Pre-entrenado			
		Clasificador	f1 score			Entrena	f1 score
Sentence Embeddings		Regresión logística	0.78	Fine-tuning características		Clasificador	0.797
TF-IDF con lematizado		Regresión logística	0.77	Fine-tuning parcial		Clasificador + 2 capas	0.824
				Fine-tuning total		Transformer + Clasificador	0.816

Recomendaciones a futuro

30

- Obtención de información más relevante (discriminatoria) que la sinopsis.
 - Agregar más datos, no solo la sinopsis, (autor, segmentos del libro como la introducción, el propio contenido del libro).
 - Interesante: explorar agregar reviews, obtención de características como conteo de adjetivos, qué adjetivos/verbos aparecen, características sintácticas...
- Definición más precisa de la etiqueta.
 - No es un problema cerrado y admite ambigüedad.
 - Seguir un estándar.
- Ante modelos con resultados similares es conveniente hacer un estudio de los requisitos computacionales y posible escalado de los modelos.
- Balanceo de las etiquetas, después del preprocesado.



¡Muchas gracias!



PROYECTO

NLP - Libros



A. C. Luis
alcaluis@alumni.uv.e
s

C. B. Rebeca
recombar@alumni.uv.e
s

M. M. Marta
memuiz@alumni.uv.e
s

V. Alejandra
ave6@uv.es

Grupo 1

