EXIST: sEXism Identification in Social neTworks

E. David Guzmán Ramírez

Licenciatura en Ciencia de Datos
Minería de Textos

Dra. Helena Gómez Adorno
M. en C. Ricardo Montalvo Lezama

9 de junio de 2021

Contenidos

Introducción

EXIST: sEXism Identification in Social neTworks es una competencia en el marco de IBERLEF 2021 (Iberian Languages Evaluation Forum) para la detección de información dañina.

De acuerdo al diccionario de Oxford, se entiende como **sexismo** *los prejuicios, estereotipos o discriminación, generalmente contra las mujeres, en base a su género*. La desigualdad y la discriminación contra las mujeres que siguen arraigadas en nuestra sociedad se reproducen cada vez más en línea.

Motivación

Detectar el sexismo online puede resultar complicado, ya que puede expresarse de formas muy diferentes. El objetivo es la detección del sexismo en un sentido amplio, desde la misoginia explícita hasta otras expresiones sutiles que involucran comportamientos sexistas implícitos.

La **identificación automática** de sexismo en un sentido amplio puede ayudar a crear, diseñar y determinar la evolución de nuevas políticas de igualdad, así como fomentar mejores comportamientos en la sociedad.

Descripción del problema

Se pedirá a los participantes que clasifiquen *tweets* y *gab*, tanto en inglés como español, de acuerdo con las dos tareas siguientes:

1. **Tarea 1: Identificación del sexismo.** Es un problema de clasificación binaria, los sistemas tienen que decidir si un texto dado es sexista o no.
2. **Tarea 2: Categorización del sexismo.** Una vez que un mensaje ha sido clasificado como sexista, la segunda tarea tiene como objetivo categorizar el mensaje según el tipo de sexismo, de acuerdo a la siguiente clasificación:
    1. Desigualdad.
    2. Estereotipos y dominio.
    3. Cosificación.
    4. Violencia sexual.
    5. Misoginia y violencia no sexual.

Análisis exploratorio

El conjunto de datos EXIST incorpora cualquier tipo de expresión sexista o fenómenos relacionados, incluidas las afirmaciones descriptivas o informadas donde el mensaje sexista es un informe o una descripción de un comportamiento sexista.

Los textos fueron extraídos de varias cuentas de Twitter y la red social Gab, estos últimos sólo están en conjunto de prueba para para medir la diferencia entre redes sociales con y sin *control de contenido*.

Datos	**Inglés**	**Español**
Train	3,436	3,541
Test	2,208	2,160

Tabla 1: Cantidad de textos en el conjunto de entrenamiento y prueba.

Análisis exploratorio

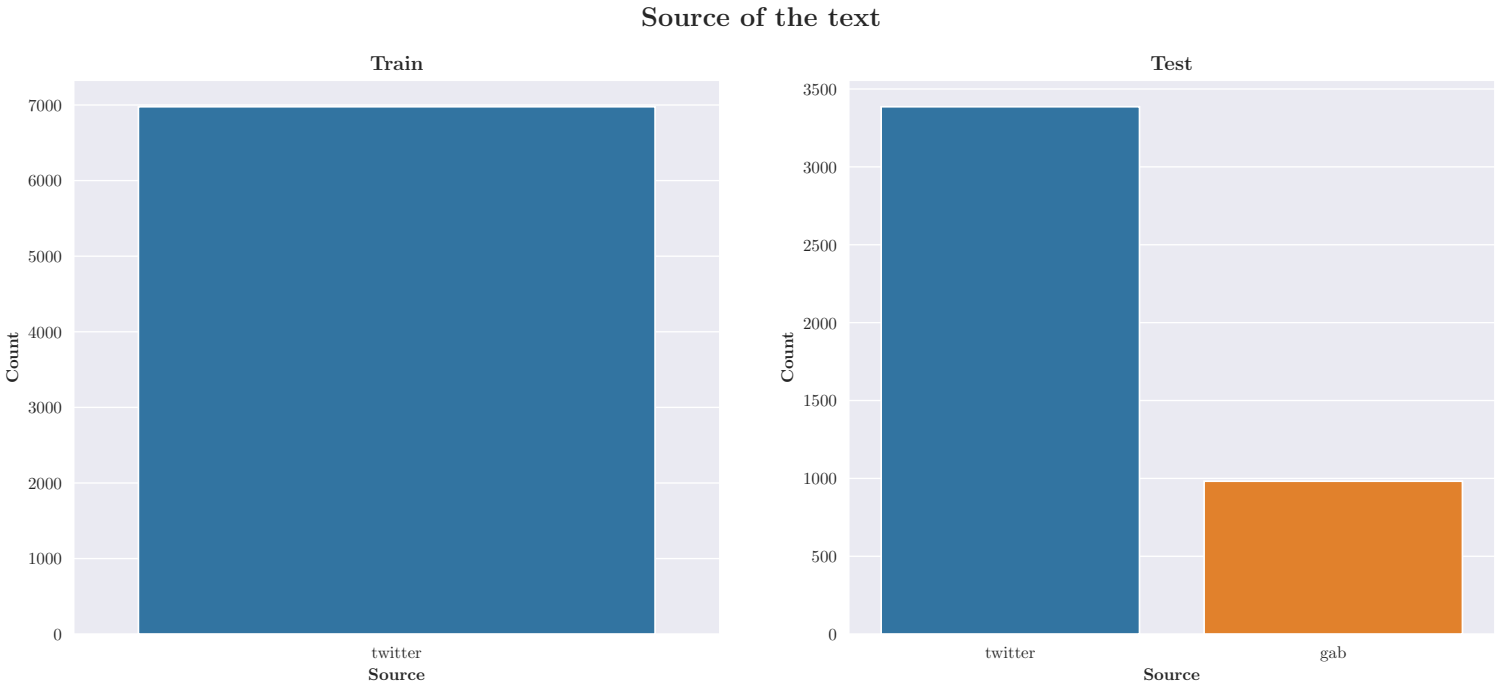

Figura 1: Procedencia de los textos en entrenamiento y prueba.

Análisis exploratorio

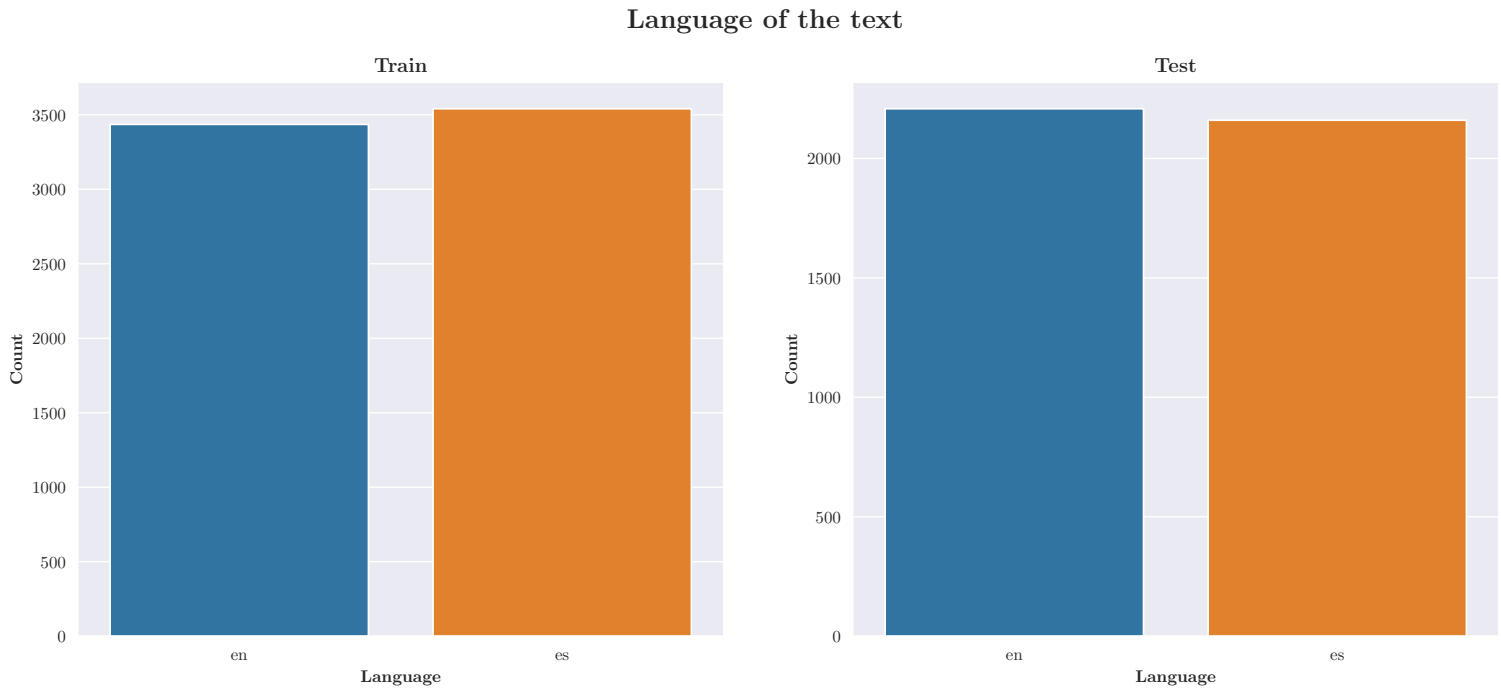

Figura 2: Idioma de los textos en entrenamiento y prueba.

Análisis exploratorio

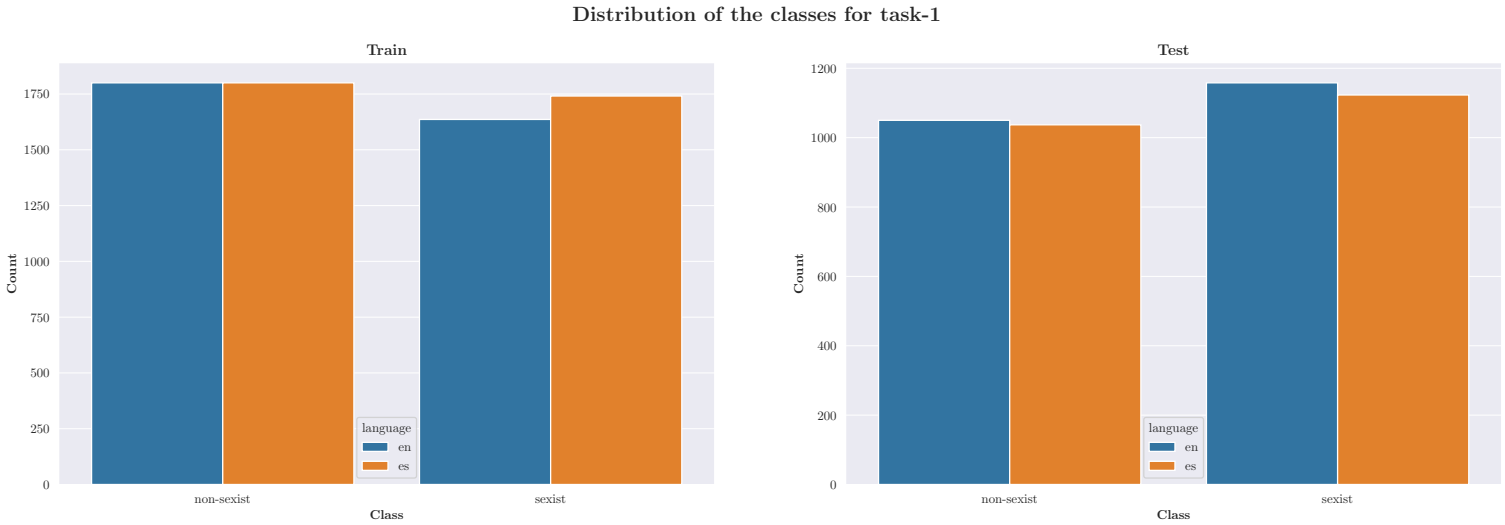

Figura 3: Distribución de las clases en entrenamiento y prueba para la primera tarea.

Análisis exploratorio

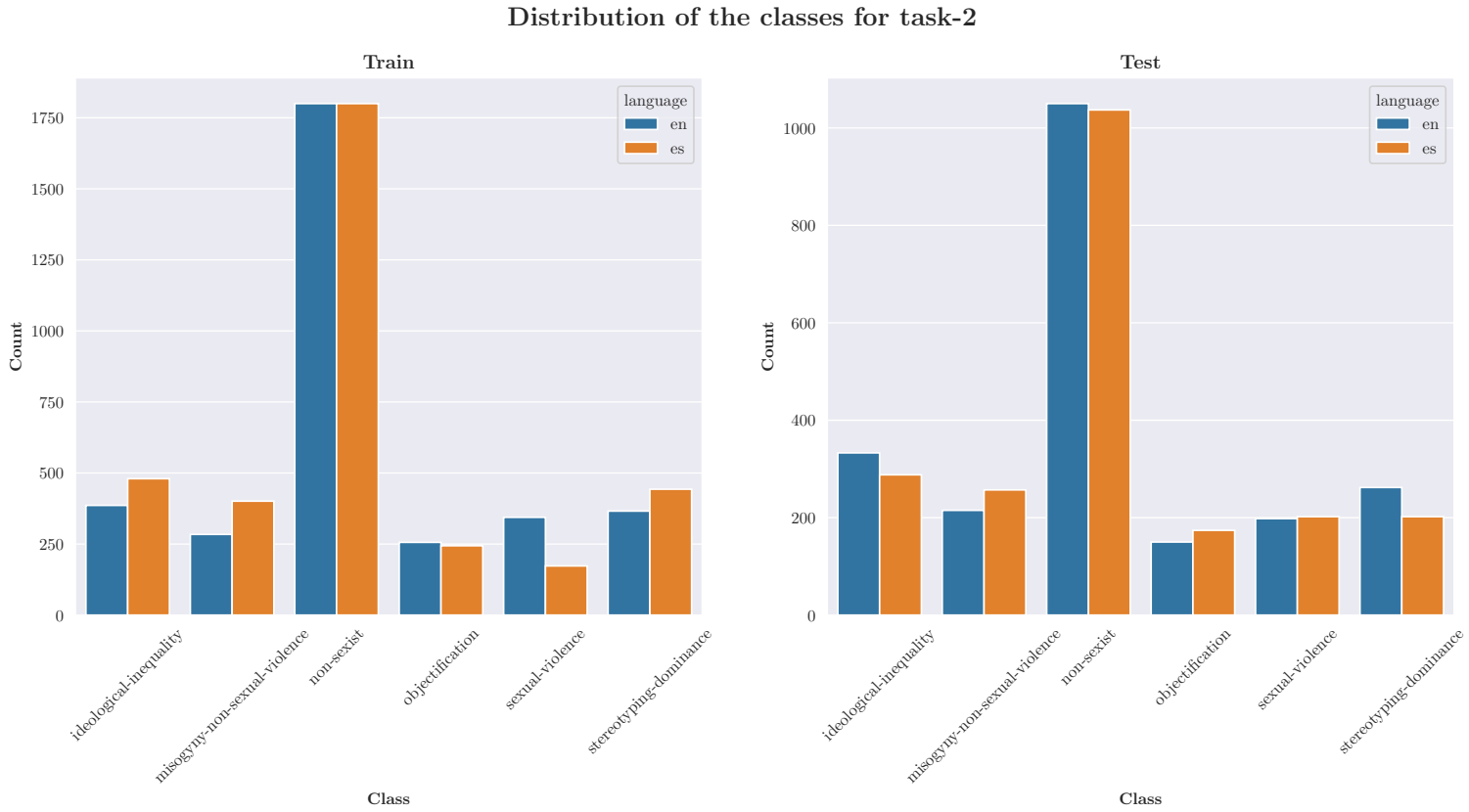

Figura 4: Distribución de las clases en entrenamiento y prueba para la segunda tarea.

Propuesta de solución

1. Se hizo un pre-procesamiento simple del texto (pasar a minúsculas, quitar nombres de usuarios y urls), intentando además aplicar un spell-checker, ya que los textos de las redes sociales suelen contener muchos errores.
2. Se usaron distintos algoritmos para sacar características del texto, haciendo la distinción entre español e inglés, y se aplicó una regresión logística para hacer la clasificación en ambas tareas.
3. Su usó un modelo de BERT (bert-base-multilingual-uncased) entrenado en varios idiomas y una red neuronal para mejorar los resultados anteriores.

Resultados tarea-1

Modelo	Task-1 (accuracy)		
	Inglés	Español	Total
tf-idf + LogisticRegression	70.70	69.81	70.26
GloVe + LogisticRegression	66.35	67.82	67.08
Doc2Vec + LogisticRegression	57.43	59.31	58.36
sentence-BERT + LogisticRegression	73.73	72.64	73.19
BERT-base-multilingual-uncased	74.86	75.69	75.27

Tabla 2: Resultados para los distintos modelos para la primer tarea.

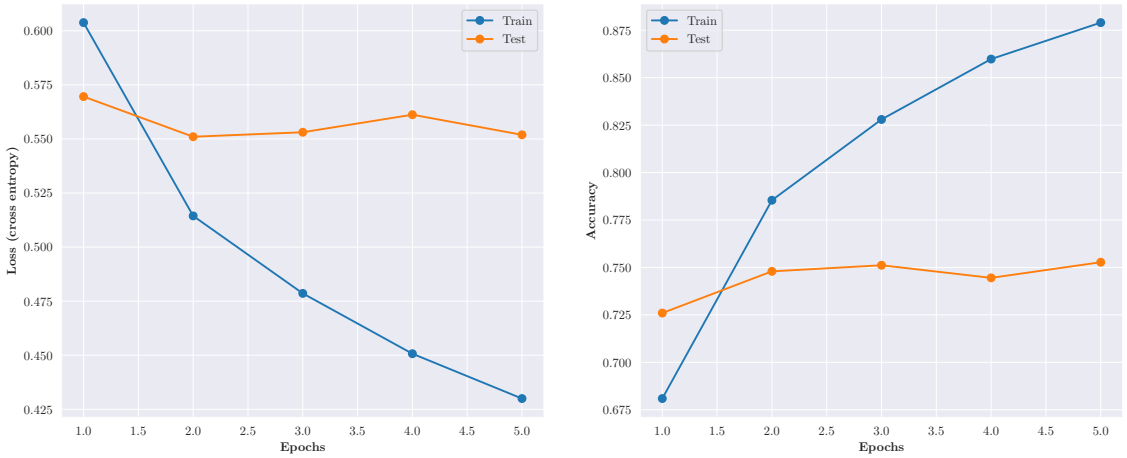

Figura 5: Fine tuning de BERT-base-multilingual-uncased.

Resultados tarea-2

Modelo	Task-2 (F1 score)		
	Inglés	Español	Total
tf-idf + LogisticRegression	54.64	50.77	53.30
GloVe + LogisticRegression	53.19	48.18	51.39
Doc2Vec + LogisticRegression	38.18	37.87	38.41
sentence-BERT + LogisticRegression	59.99	61.08	60.76
BERT-base-multilingual-uncased	61.67	62.95	62.38

Tabla 3: Resultados para los distintos modelos para la segunda tarea.

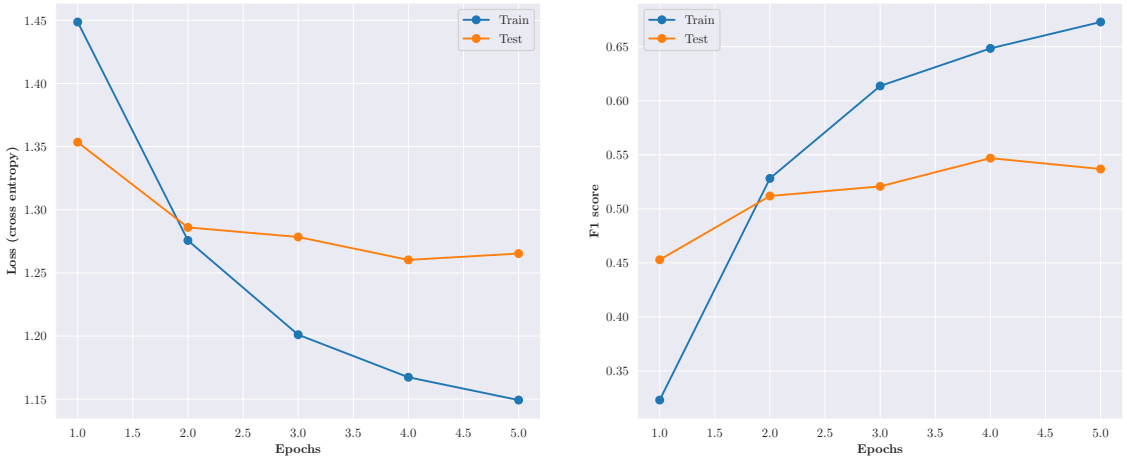

Figura 6: Fine tuning de BERT-base-multilingual-uncased.

Conclusiones

1. Observando los mejores resultados de la competencia, 78.04 de accuracy para la tarea 1 y 57.87 de F1 para la tarea 2, los modelos propuestos obtienen resultados sumamente cercanos, especialmente el que usa BERT-base-multilingual-uncased (75.27 y 62.38 respectivamente).
2. El spell-check debería funcionar para mejorar el pre-procesamiento, pero se debería usar un diccionario con el *slang* de cada idioma para que de mejores resultados.