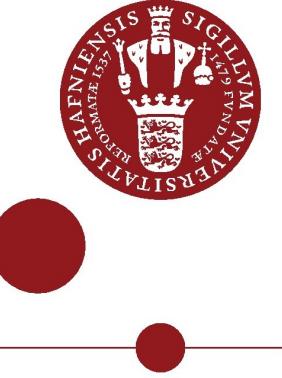




UNIVERSITAT DE
BARCELONA

UNIVERSITY OF
COPENHAGEN



MuSeD: corpus multimodal en español para la detección de sexismo en vídeos de redes sociales

Laura De Grazia, Pol Pastells, Mauro Vázquez Chas, Desmond Elliott,

Danae Sánchez Villegas, Mireia Farrús, Mariona Taulé

arXiv preprint arXiv:2504.11169

lauradegrazia@ub.edu

Mayo 2025

Índice

1. Introducción

2. Motivación

3. Objetivos

4. MuSeD: creación y anotación

5. Análisis del corpus

6. Detección automática del sexismo

7. Análisis de la detección automática

8. Conclusiones



WARNING

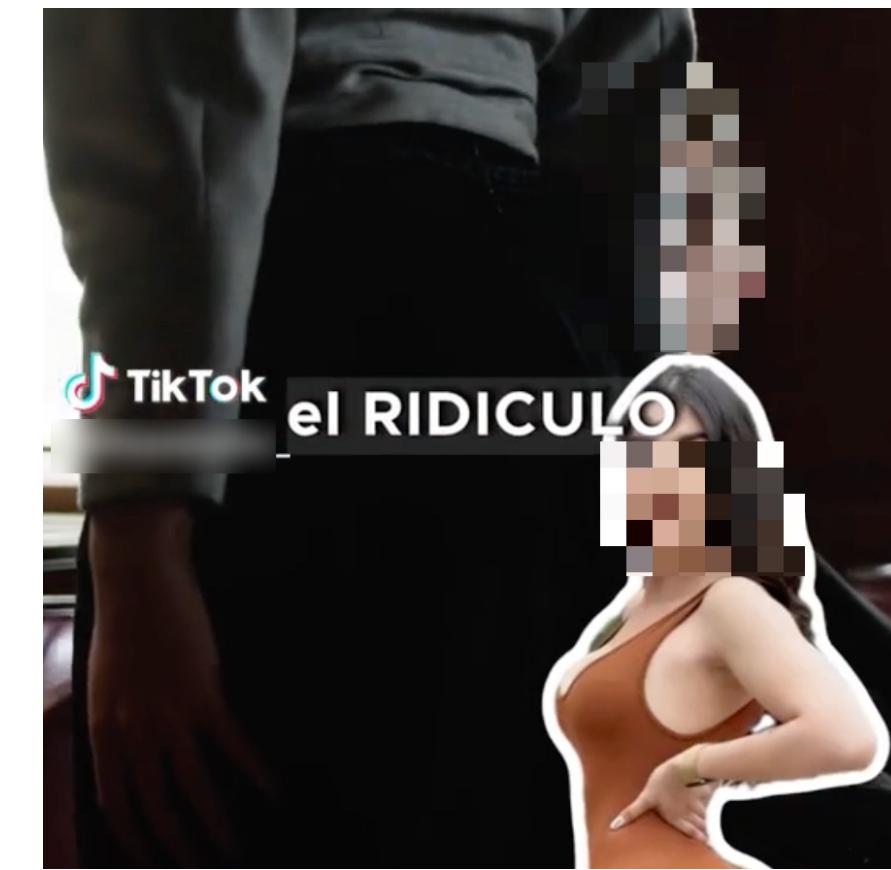
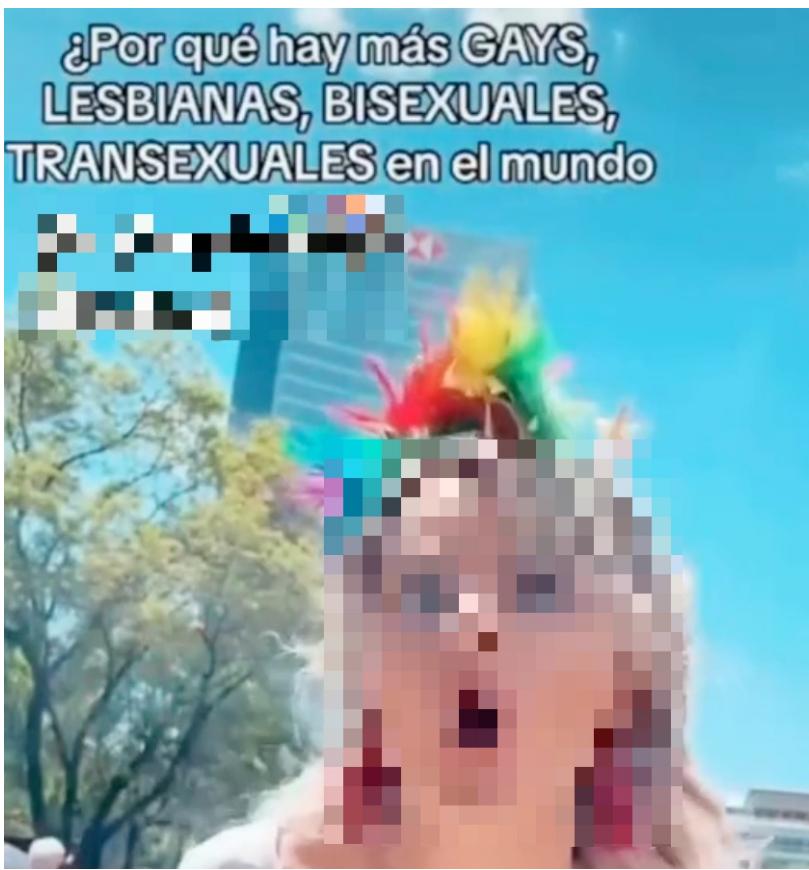
Esta presentación contiene ejemplos de lenguaje
e imágenes que pueden resultar ofensivos.

No representan mi punto de vista.



Introducción

¿Cuándo se considera sexista un contenido?



Estereotipo

Mi gente, mi gente, miren, disfruten de estos juegos que traen plancha, lavadora, que traen así, porque las feministas pronto van a protestar que las niñas no deben jugar con esto.

Negación de desigualdad

Los hombres tenemos privilegios por sobre a las mujeres, sociales y legales. Tenemos seis años más de pena por el mismo delito cometido por una mujer y el doble de posibilidades de ser encarcelado.

Discriminación

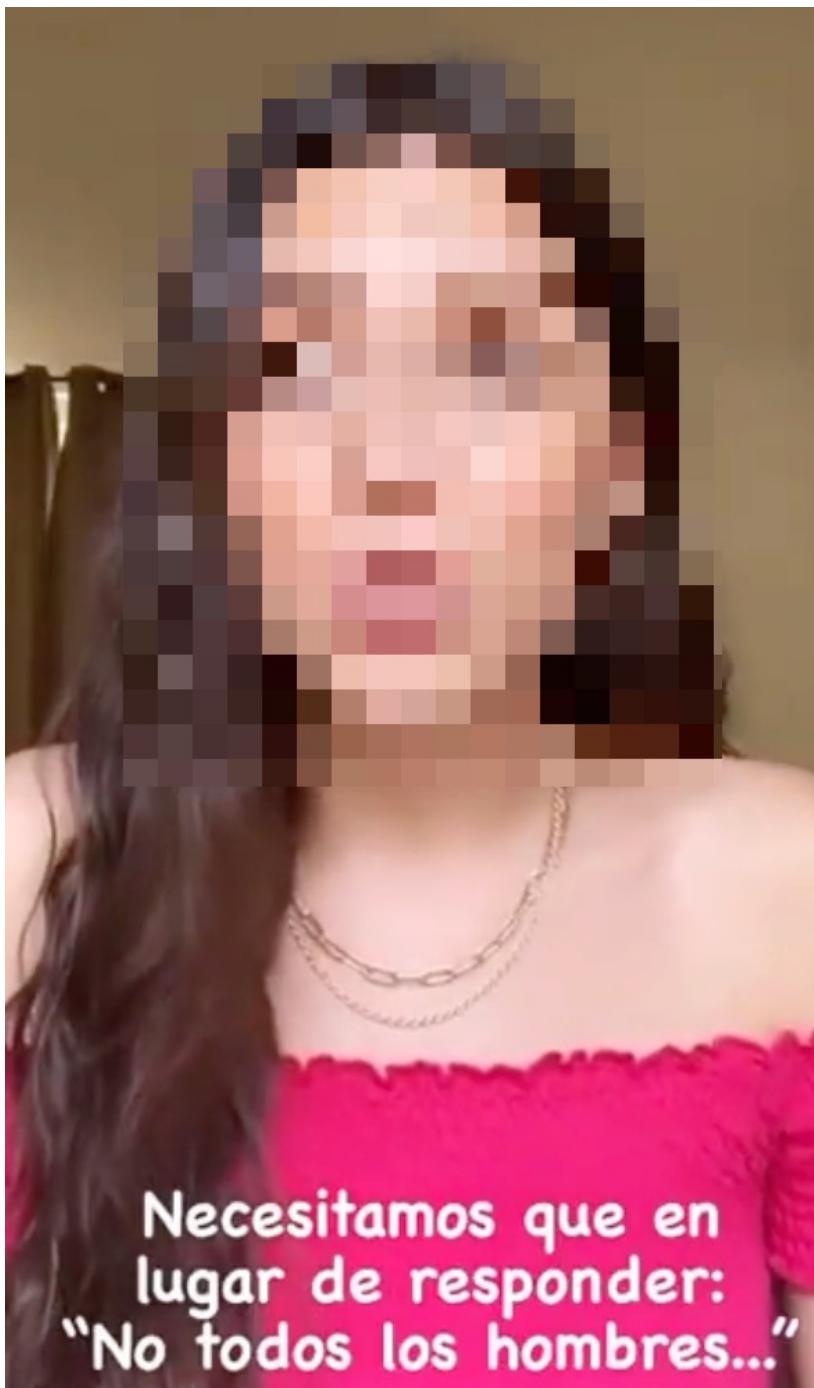
¿Por qué hay más gays, lesbianas, bisexuales, transexuales en el mundo? Porque les vendieron la mentira de la identidad sexual.

Objetivación

¿Cómo ser gracioso y conquistar sin hacer el ridículo? Mi hermano, una habilidad clave es la malinterpretación intencional.

MuSeD: anotación del corpus

¿Cuándo se considera no-sexista un contenido?



Necesitamos que en lugar de responder no todos los hombres para defender que son la excepción, mejor nos escuchen.

Formaremos parte de la solución cuando entendamos qué es el privilegio masculino y cuestionemos qué es la violencia machista y la masculinidad tóxica.

Motivación

La detección automática del sexismio en las plataformas sociales es importante.

- Apoya el trabajo de los moderadores humanos (Zeinert et al., 2021)
- Ayuda a prevenir la difusión de narrativas tóxicas

BBC

Facebook and YouTube moderators sign PTSD disclosure



Content moderators are being asked to sign forms stating they understand the job could cause post-traumatic stress disorder (PTSD), according to reports.

Motivación

Enfoque limitado en los estudios sobre el sexismo multimodal.

- La mayoría de los estudios se enfocan sobre la detección de la misoginia y del sexismo en los **memes** (Fersini et al., 2022; Plaza et al., 2024)
- El único estudio previo sobre la detección del sexismo en los videos es *Sexism identification on TikTok* (Arcos et al., 2024)
- Poca consideración del sexismo hacia identidades no cisgénero y no heterosexuales



(MAMI dataset, Fersini et al., 2022)

Motivación

La detección automática del sexismoy en los videos es **difícil**.

- En el video tenemos que considerar **más modalidades** (texto, audio, imagen)
- Casos en los que hay una contradicción entre el contenido verbal y el contenido no verbal



Soy una persona trans, no binaria. Más específicamente soy demigirl. Mi orientación sexual es polisexual.

Motivación

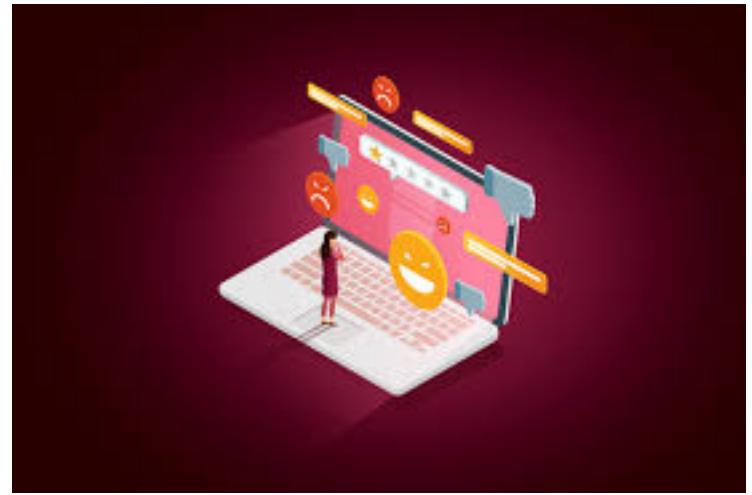
Casos difíciles de clasificar por factores que dependen de la interpretación subjetiva.

- **Sexismo implícito** (contexto cultural y social)
- Uso de **ironía y sarcasmo**



Los hombres tenemos privilegios
por sobre a las mujeres, sociales y legales.
Tenemos seis años más de pena
por el mismo delito cometido
por una mujer y el doble
de posibilidades de ser encarcelado.

Objetivos



Introducción de MuSeD (Multimodal Dataset for Sexism Detection)



Nuevo sistema de anotación



Evaluación de LLMs y LLMs multimodales en la detección de sexismo utilizando MuSeD

MuSeD: fuentes de datos



Plataforma con moderación de contenidos :

- **Prohibición de *hate speech***, incluidos discrimination por sexo, orientación sexual y identidad de genero (TikTok Community guidelines, 2024)
- **Persisten formas de denigración** tanto de manera explícita como implícita (Banet-Weiser et al., 2023)

Plataforma con baja moderación de contenidos:

- **No se previene la difusión de *hate speech***
- **Difusión de *hate speech*** por usuarios expulsados de otras plataformas y por quienes aún mantienen perfiles en redes sociales convencionales (Trujillo et al., 2020)

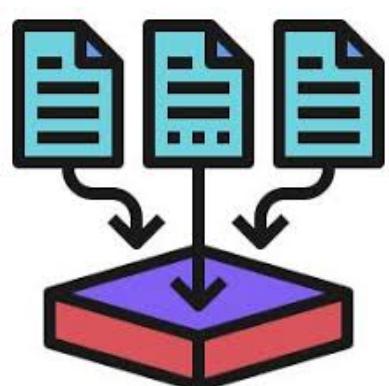
MuSeD: recopilación & preprocesamiento



- Uso de 187 hashtags (#estereotipos de género, #feminismo, #ideología de género [...])
- Identificación de usuarios que comparten contenidos sexistas (Anzovino et al., 2018)



- Adaptación de una muestra de hashtags para usarlos como keywords



Recopilación de videos

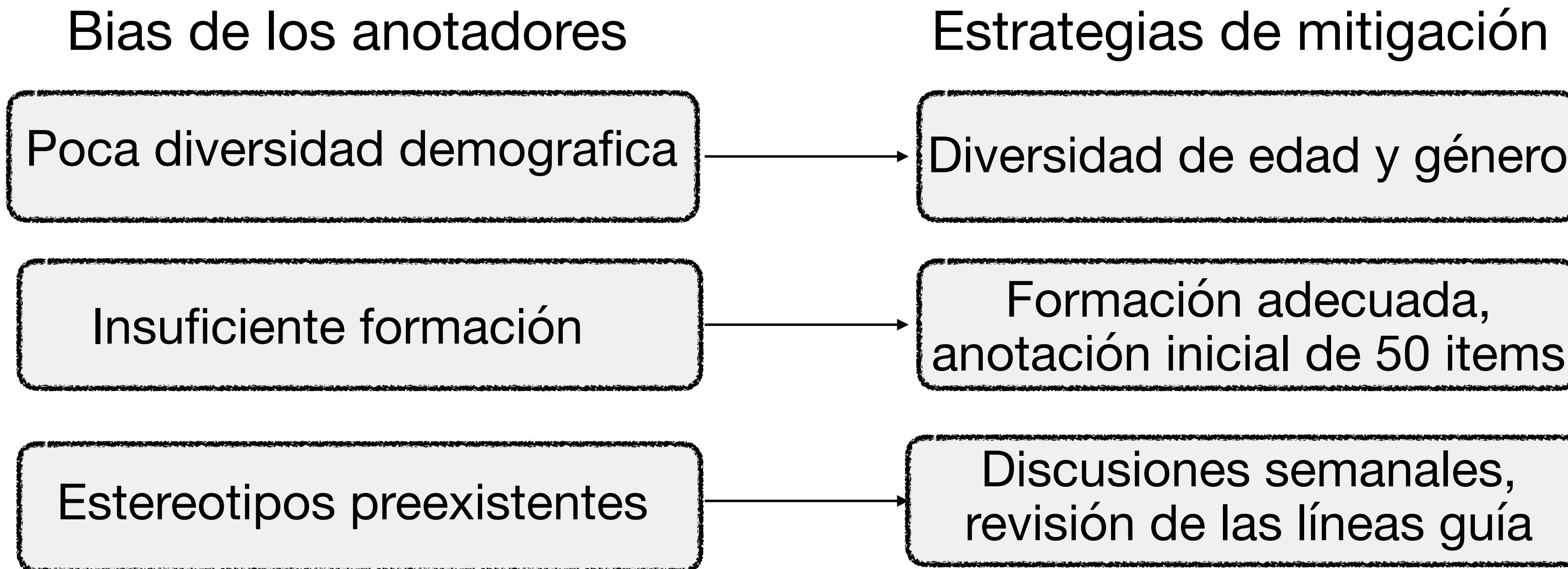
- TikTok: Apify (Arcos et al., 2024)
- BitChute: BitChute dl software (Das et al., 2023)



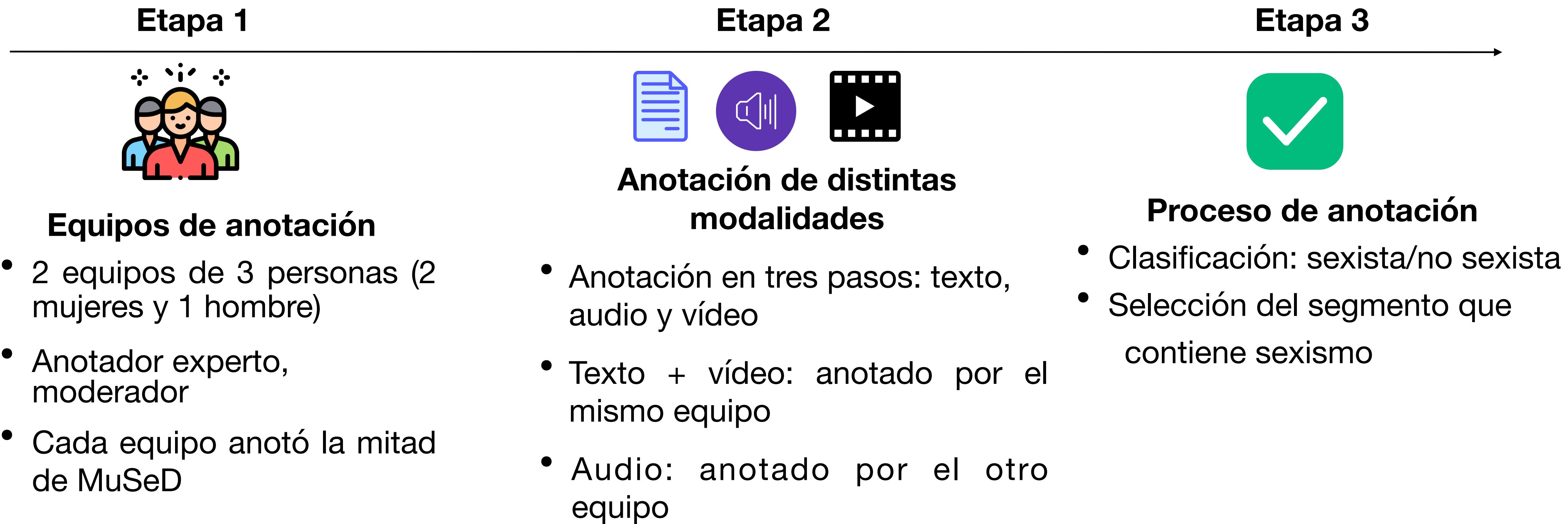
Preprocesamiento

- Texto: Whisper-CTranslate2
- Audio: línea de comando FFmpeg
- OCR: Python module EasyOCR

MuSeD: anotación del corpus



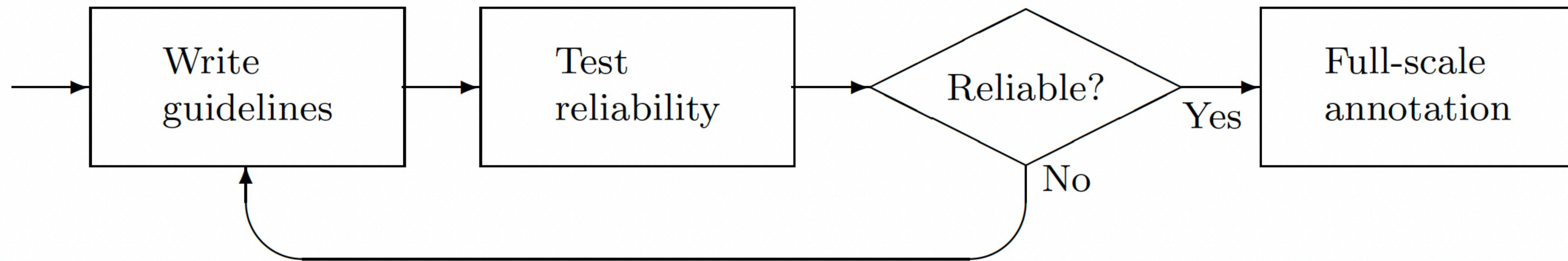
MuSeD: anotación del corpus



MuSeD: Inter-Annotator Agreement

¿Cómo podemos comparar los resultados de las anotaciones realizadas por distintos anotadores?

- Para compararlos, es necesario calcular el **Inter-Annotator Agreement (IAA)**.
- El resultado del IAA indica si el proceso de anotación es **fiable y reproducible**.
- Antes de anotar el entero corpus, es buena práctica conducir una prueba de acuerdo (*agreement testing*).



Arstein, 2017

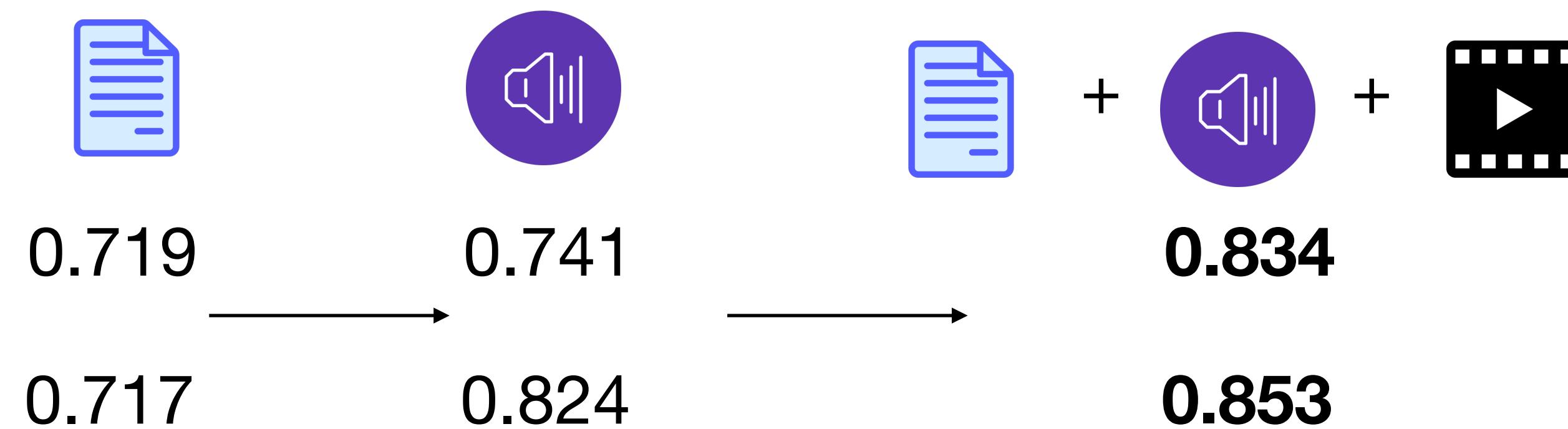
MuSeD: Inter-Annotator Agreement

Para evaluar el IAA, hemos utilizado el coeficiente **Fleiss' Kappa**.

- Se aplica cuando hay **más de dos anotadores**
- Se emplea para medir el IAA más allá del nivel que se esperaría por azar o por una codificación arbitraria.
- El valor de Kappa varía entre -1 y 1:
 - 1 indica acuerdo perfecto
 - 0 indica un acuerdo equivalente al que se esperaría por azar
 - valores negativos reflejan menos acuerdo del esperado por azar

MuSeD: Inter-Annotator Agreement

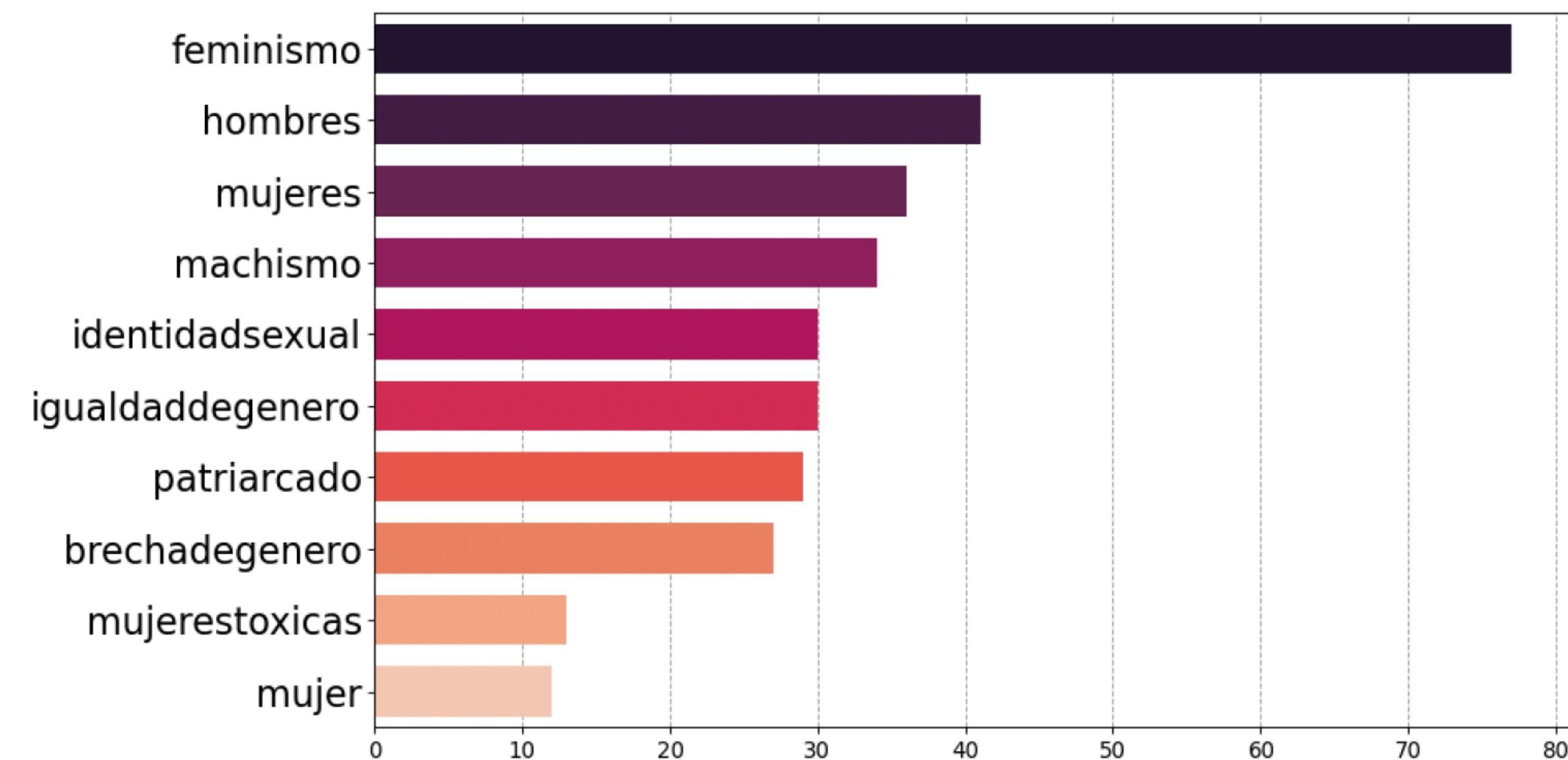
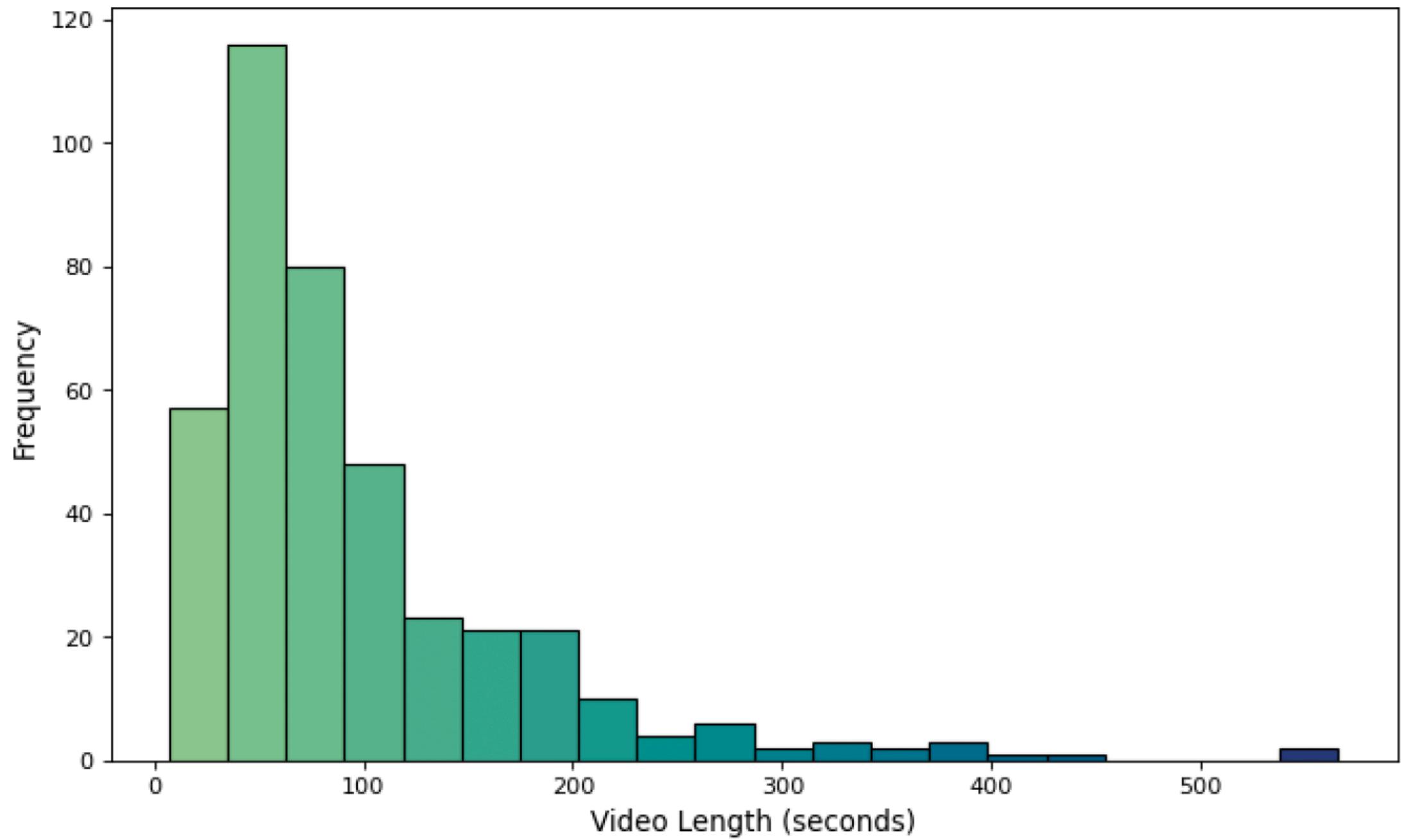
- Cálculo del Fleiss' Kappa para cada modalidad (texto, audio) y combinación de modalidades (vídeo)
- Comparación del Fleiss' Kappa entre distintas modalidades
- Aumento del Fleiss' Kappa de *substancial* (0.61 – 0.80) a *almost perfect* (0.81 – 1.00) al considerar todas las modalidades



- IAA superior al de Arcos et al. (2024). $\kappa = 0.499$ (moderate agreement)

MuSeD: análisis del corpus

- 400 videos, \approx 11 horas, (\approx 90% TikTok, \approx 10%BitChute)
- 48,5 % de los videos etiquetados como sexistas y 51,5 % como no-sexistas



Detección automática del sexismo

Evaluación de una variedad de LLMs en la tarea de detección de sexismo usando MuSeD.

- Modelos evaluados en **configuración zero-shot**
- Entrada: solo texto o texto + imagen
- **Clasificación binaria:** sexista / no sexista
- Se analiza el impacto de la **entrada multimodal** en el rendimiento
- Para evaluar el rendimiento, hemos utilizado el accuracy

Detección automática del sexism

System prompt

Eres un asistente que sigue patrones y solo puede responder con 'Sí' o 'No'. Tu objetivo es determinar si un texto es sexista.

User prompt

Instrucción:

Determina si el texto es sexista.

Texto:

[...]

Es el texto sexista?:

Prompt utilizado para clasificar las transcripciones de texto de vídeos usando modelos solo de texto

Detección automática del sexismo

System prompt

Eres un asistente de clasificación de videos. Responde solo con 'Sí' o 'No'.

User prompt

Instrucción:

Analiza el contenido de este video y determina si es sexista. Si el texto del video es sexista, el video también debe clasificarse como sexista, incluso si las imágenes no parecen ofensivas.

Texto:

[...]

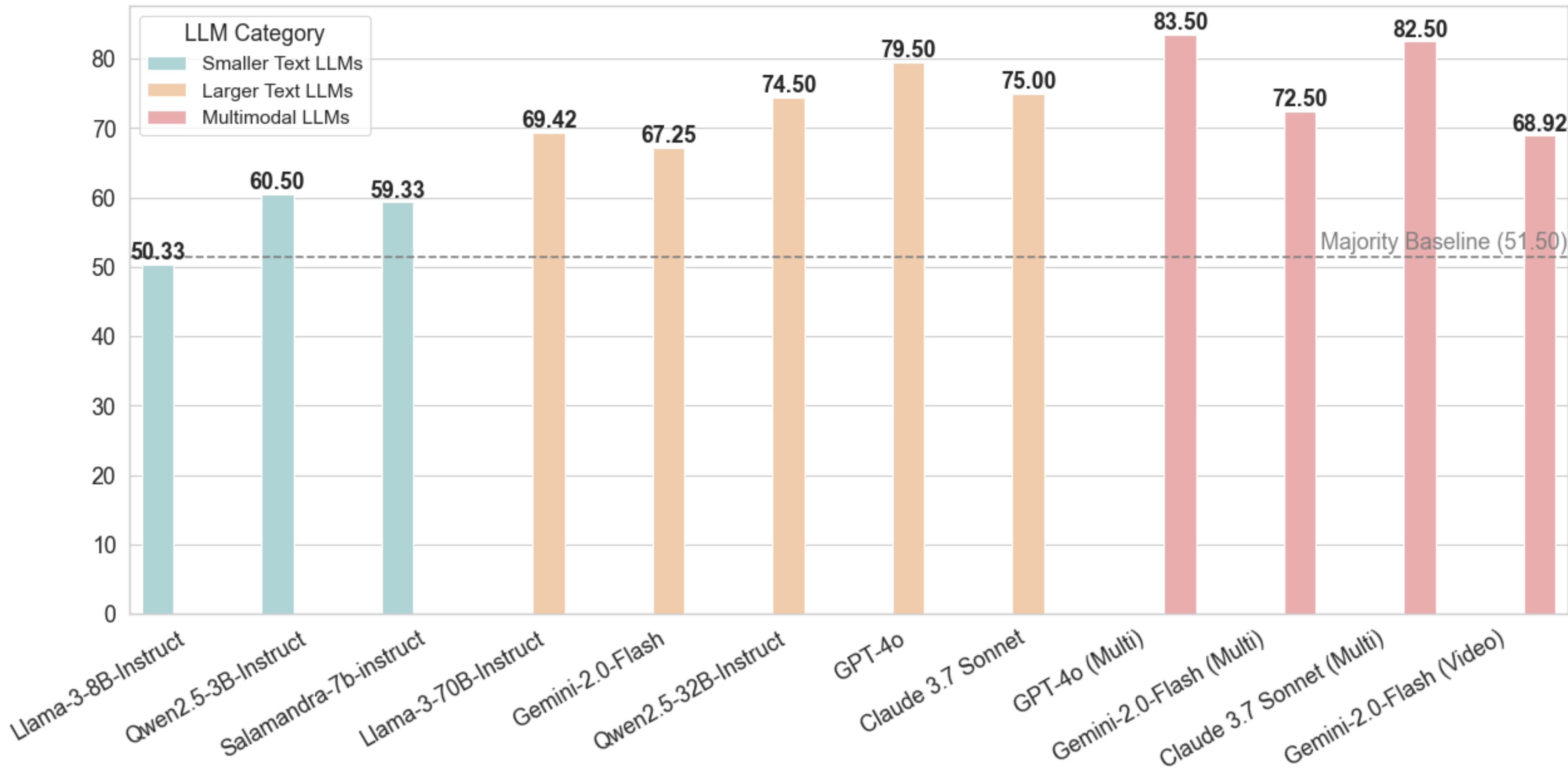
Pregunta:

Basado en el texto y las imágenes proporcionadas, ¿el video es sexista? Responde solo con 'Sí' o 'No'. Si el texto es sexista, responde 'Sí' sin importar las imágenes.

Prompt utilizado para clasificar videos usando LLMs multimodales

Análisis

Incremento del rendimiento de **GPT-4o** al incluir **entrada multimodal** (texto + imagen)



Análisis

Análisis de errores de GPT-4o

- **La incorporación de contenido visual reduce los errores:**
 - 82 en la configuración solo texto y 72 en texto + imagen

⚠ Persisten 44 errores en ambas configuraciones → indican desafíos inherentes en ciertos casos de sexismo

Análisis

- 20% de los errores de GPT-4o corresponden a casos de **sexismo implícito**
- Casos también difíciles para los anotadores humanos
- Disminución del IAA en estos casos:
 - IAA en sexismio implícito: 0.61/0.70
 - IAA en el dataset: 0.83/0.85
- Difíciles de detectar por **factores sociales y culturales** que influyen en la interpretación



Alpha. Piernas separadas. Cruzado masculino. El frente.
Inclinarse hacia atrás.
La cabeza. Brazos cómodos.

Conclusiones

- **Nuevo dataset multimodal** con vídeos etiquetados como sexista y no sexista
- Primer dataset con sistema de **anotación con distintas modalidades (texto, audio y video)**, usando **TikTok y BitChute** como plataformas y con definición ampliada del sexismo
- **Experimentos de evaluación** usando texto, combinación de texto y contenido visual para la detección de sexismo
- **Enfoque multimodal** mejora la clasificación del sexismo tanto por anotadores humanos como por modelos
- **Próximos objetivos:**
 - expansión del dataset
 - clasificación mas detallada de distintos tipos de sexismo

Gracias



De Grazia, L., Pastells, P., Chas, M. V., Elliott, D., Villegas, D. S., Farrús, M., & Taulé, M. (2025). MuSeD: A Multimodal Spanish Dataset for Sexism Detection in Social Media Videos. arXiv preprint arXiv:2504.11169.