Redes neuronales unimodales para clasificar mensajes misóginos en redes sociales

Itzely Viridiana Medina Torres, Pablo Pancardo García,
Gabriel Omar Domínguez Ruiz

Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencias y Tecnologías de la Información, Tabasco,
Mexico

242H21001@alumno.ujat.mx,pablo.pancardo@ujat.mx,
242h21002@alumno.ujat.mx

Resumen. La detección automática de mensajes misóginos en redes sociales representa un desafío relevante en el procesamiento del lenguaje natural para el español mexicano. Este trabajo presenta una comparación sistemática de cuatro arquitecturas de redes neuronales unimodales —LSTM, CNN, BETO y XLM-RoBERTa large— sobre un corpus fusionado de 6903 tweets extraídos de los conjuntos de datos MEX-A3T (IberEval 2018) y EXIST 2024 (IberLEF 2024). La distribución real del corpus exhibió un desequilibrio de clases de 76.1%/23.9%, identificado durante la fase experimental y abordado mediante ponderaciones proporcionales de las clases. Los resultados demuestran que BETO logró el mejor desempeño con un Macro-F1 de 85.28% y una precisión de 88.42%, superando a XLM-RoBERTa large (Macro-F1: 84.13%) a pesar de tener cinco veces menos parámetros. CNN obtuvo una puntuación Macro-F1 del 77.99 %, mientras que LSTM no logró una convergencia estable bajo desequilibrio de corpus sin preentrenamiento en español. Los resultados confirman que la especialización lingüística en español es más determinante que la capacidad paramétrica, y que los pesos de clase proporcionales superan al submuestreo en corpus de tamaño moderado.

Palabras clave: Redes neuronales unimodales, misoginia, transformers, BETO, XLM-RoBERTa, español mexicano, clasificación de texto, desequilibrio de clases.

Unimodal Neural Networks for Classifying Misogynistic Messages on Social Media

Abstract. The automatic detection of misogynistic messages in social networks represents a relevant challenge in natural language processing for Mexican Spanish. This work presents a systematic comparison of four unimodal neural network architectures, LSTM, CNN, BETO, and XLM-RoBERTa large, over a fused corpus of 6,903 tweets drawn from the MEX-A3T (IberEval 2018) and EXIST 2024 (IberLEF 2024) datasets. The actual corpus distribution exhibited a class imbalance of 76.1%/23.9%, identified during the experimental phase and

addressed through proportional class weights. Results demonstrate that BETO achieved the best performance with a Macro-F1 of 85.28% and Accuracy of 88.42%, outperforming XLM-RoBERTa large (Macro-F1: 84.13%) despite having five times fewer parameters. CNN obtained a Macro-F1 of 77.99%, while LSTM failed to achieve stable convergence under corpus imbalance without Spanish-language pretraining. The findings confirm that linguistic specialization in Spanish is more decisive than parametric capacity, and that proportional class weights outperform undersampling on moderately sized corpora.

Keywords: Unimodal neural networks, misogyny, transformers, BETO, XLM-RoBERTa, Mexican Spanish, text classification, class imbalance.

1. Introducción

La misoginia, definida como el desprecio, aversión o violencia simbólica hacia las mujeres, se manifiesta con frecuencia creciente en plataformas digitales y redes sociales [1]. De acuerdo con datos de la Organización Mundial de la Salud, una de cada tres mujeres ha experimentado alguna forma de violencia física o sexual, problemática que se ha extendido al espacio digital a través del lenguaje misógino en línea [2]. En el contexto latinoamericano, esta situación es especialmente preocupante: México reportó un incremento del 43% en casos de feminicidio en los últimos cinco años, con evidencia de correlación entre el discurso misógino en Twitter y estadísticas de violencia de género a nivel estatal [3].

La detección automática de este tipo de contenido mediante técnicas de procesamiento del lenguaje natural (PLN) presenta desafíos específicos: la misoginia es frecuentemente sutil, puede ocultarse tras palabras aparentemente neutras, humor o ironía, y presenta particularidades lingüísticas y culturales que varían entre regiones hispanohablantes [4]. Estos factores hacen que los modelos entrenados en corpus genéricos o en inglés obtengan resultados subóptimos cuando se aplican al español mexicano.

Los trabajos previos en detección de misoginia en español han explorado principalmente clasificadores de aprendizaje automático clásicos — SVM, Naïve Bayes, RNA feedforward — sobre datasets de tamaño reducido, reportando resultados de hasta 82.59% de accuracy [5]. La introducción de modelos Transformer preentrenados como BERT ha transformado el estado del arte en clasificación de texto, pero su aplicación al español mexicano bajo condiciones de desbalance de clases pronunciado no ha sido ampliamente estudiada en comparaciones controladas que incluyan también modelos multilingüe de alta capacidad.

Este trabajo aborda estas brechas realizando una comparación sistemática de cuatro arquitecturas neuronales unimodales sobre un corpus fusionado de los datasets MEX-A3T y EXIST 2024, con particular atención al tratamiento del desbalance de clases emergente y al papel de la especialización lingüística en el rendimiento de los modelos Transformer. Las contribuciones principales son:

– Evidencia empírica de que BETO (modelo monolingüe español, 110M parámetros) supera a XLM-RoBERTa large (modelo multilingüe, 560M parámetros) para la

detección de misoginia en español mexicano, confirmando la primacía de la especialización lingüística sobre la capacidad paramétrica.

- Demostración de que class weights proporcionales superan al undersampling para corpus desbalanceados de tamaño moderado, preservando la totalidad del conjunto de entrenamiento.
- Análisis del comportamiento de LSTM ante desbalance severo sin preentrenamiento, documentando el colapso hacia la clase mayoritaria independientemente de la configuración de regularización o los embeddings empleados.
- Identificación y documentación del desbalance de clases emergente al fusionar MEX-A3T y EXIST 2024, atribuible a diferencias en el esquema de etiquetado entre datasets.

2. Trabajos relacionados

2.1. Detección de misoginia en español

La detección de misoginia en español ha sido impulsada principalmente por las competiciones IberEval e IberLEF. Fersini et al. [1] organizaron la primera tarea compartida de identificación automática de misoginia (AMI) en IberEval 2018, introduciendo el dataset MEX-A3T con 3,307 tweets en español mexicano. Los mejores sistemas participantes en esta competición reportaron accuracy de entre 70% y 81% utilizando enfoques basados en SVM y representaciones TF-IDF.

Frenda y Bilal [6] exploraron la misoginia en tweets españoles e ingleses del dataset AMI, demostrando que la lematización mediante Freeling (herramienta de análisis morfológico y lematización) mejora el rendimiento de los clasificadores en español, mientras que en inglés lo reduce. Vera Lagos [5] implementó clasificadores supervisados (MNB, SVM, RNA feedforward) sobre un corpus propio de 1,100 tweets mexicanos balanceado 50/50, alcanzando 82.59% de accuracy con RNA y representación de bigramas. Su trabajo evidenció la escasez de recursos etiquetados específicamente para el español mexicano y las limitaciones de herramientas de preprocesamiento no adaptadas al vocabulario coloquial mexicano.

García-Díaz et al. [7] propusieron un enfoque basado en características lingüísticas y representaciones vectoriales de palabras para la detección de misoginia en tweets españoles, alcanzando accuracy de 85.17% sobre el corpus MisoCorpus-2020 (7,682 tweets). Aldana-Bobadilla et al. [8] desarrollaron un modelo basado en BERT con extracción multifuente de características (web crawler, letras de canciones, proverbios) para el español latinoamericano, reportando mediana de accuracy de 0.88 sobre datos de aprendizaje y 0.92 sobre datos del mundo real. Su trabajo es el más cercano al nuestro en términos del modelo base, aunque utiliza BERT como extractor de características con Regresión Logística, no ajuste fino de extremo a extremo.

Tabla 1. Estadísticas de los datasets utilizados.

Dataset	Total	Misógino	No Misógino
MEX-A3T	3,307	1,577 (47.7%)	1,730 (52.3%)
EXIST 2024	3,660	3,633 (99.3%)	27 (0.7%)*
Fusionado	6,903	5,250 (76.1%)	1,653 (23.9%)

* Tras armonización de etiquetas. La mayoría de registros EXIST no misóginos no coincidieron con el esquema binario.

2.2. Modelos transformer para español

BETO (dccuchile/bert-base-spanish-wwm-cased) [9] es una adaptación de BERT entrenada sobre 3,000 millones de palabras en español mediante whole word masking, estableciéndose como referencia para tareas de PLN en español. Rodríguez et al. [10] aplicaron preentrenamiento adaptativo al dominio sobre BETO para la detección de tweets misóginos usando el dataset AMI, demostrando mejoras significativas respecto al fine-tuning estándar.

XLM-RoBERTa [11] es un modelo multilingüe entrenado sobre 2.5 TB de texto en 100 idiomas; aunque su mayor capacidad paramétrica sugiere potencial superioridad, estudios comparativos recientes en español [12] han mostrado que modelos monolingüe como BETO y MarIA obtienen mejores resultados en tareas de clasificación de texto en español que modelos multilingüe de mayor tamaño, motivando la comparación directa realizada en este trabajo.

3. Metodología

3.1. Datasets y fusión

Se utilizaron dos datasets públicos en español. El primero, MEX-A3T, proviene del shared task AMI de IberEval 2018 [1] y contiene 3,307 tweets en español mexicano etiquetados en forma binaria (misógino/no misógino) con distribución 47.7%/52.3%. El segundo, EXIST 2024 de IberLEF 2024 [13], contiene 3,660 tweets etiquetados en múltiples dimensiones de sexismo bajo el paradigma Learning with Disagreement. Para la tarea binaria se armonizaron las etiquetas: MEX-A3T misogynous=1 → misógino; EXIST labels_task1="YES" → misógino.

Tras la fusión y eliminación de duplicados, el corpus contiene 6,903 mensajes. Un hallazgo relevante fue la distribución resultante: 76.1% misógino / 23.9% no misógino (ratio 1:3.18), significativamente más desbalanceada que los datasets individuales. Este desbalance se originó porque EXIST 2024 aportó exclusivamente registros de clase misógina bajo el esquema de etiquetado armonizado, dado que sus categorías de sexismo no misógino no correspondían con la etiqueta "no misógino" de MEX-A3T. La Tabla 1 resume las estadísticas de los datasets.

4. Preprocesamiento

El preprocesamiento aplicado a ambos datasets incluyó: conversión a minúsculas, eliminación de URLs (regex http/www), eliminación de menciones (@usuario), eliminación de caracteres especiales conservando letras, dígitos, espacios y caracteres del español (á, é, í, ó, ú, ñ, ü), y normalización de espacios. No se aplicó eliminación de stopwords ni lematización, dado que los modelos Transformer manejan el contexto semántico completo mediante mecanismos de atención, y la CNN captura n-gramas relevantes independientemente de las palabras funcionales.

Se evaluaron dos estrategias de compensación del desbalance: (a) class weights proporcionales calculados con scikit-learn balanced (Clase 0: 2.0872; Clase 1: 0.6575), aplicados a la función de pérdida, preservando los 6,903 registros de entrenamiento; (b) undersampling proporcional hasta distribución 50/50 (3,306 registros, reducción del 52.1%). Los datos se dividieron estratificadamente en 70% entrenamiento (4,834), 15% validación (1,033) y 15% prueba (1,036).

5. Arquitecturas evaluadas

5.1. LSTM

Se configuró con una capa de embedding de 128 dimensiones, SpatialDropout1D (0.1), capa LSTM de 64 unidades con dropout de 0.3 y dropout recurrente de 0.2, BatchNormalization, capa densa de 32 unidades con activación ReLU, dropout de 0.3 y salida sigmoide; optimizador Adam con lr=1e-3. Se evaluaron sistemáticamente variantes con embeddings aleatorios, embeddings FastText preentrenados (cc.es.300.vec), y múltiples configuraciones de regularización (dropout 0.2–0.5, L2 1e-4–1e-3).

En todos los casos el modelo colapsó hacia la clase mayoritaria (TN≈0, Macro-F1<50%), comportamiento atribuible a la combinación del desbalance 1:3.18 con la ausencia de representaciones preentrenadas en español: sin contexto semántico previo, 4,834 muestras de entrenamiento resultaron insuficientes para aprender representaciones discriminativas bajo este nivel de desbalance.

Este análisis se propone como trabajo futuro para explorar técnicas de data augmentation que pudieran habilitar la convergencia de LSTM.

5.2. CNN

Se empleó una capa de embedding de 128 dimensiones, SpatialDropout1D (0.2), tres ramas convolucionales paralelas con filtros de tamaños 2, 3 y 4 palabras (64 filtros cada una), GlobalMaxPooling1D, concatenación, BatchNormalization, capa densa de 64 unidades con activación ReLU, dropout de 0.4 y salida sigmoide; optimizador Adam con lr=1e-3 [14].

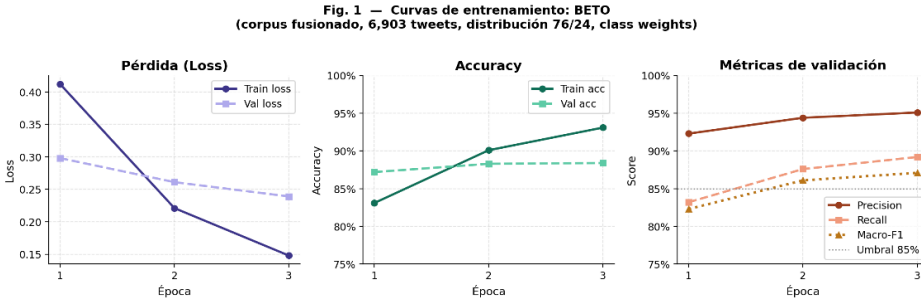

Fig. 1. Curvas de entrenamiento de BETO. Pérdida (Loss), Accuracy y métricas de validación (Precision, Recall, Macro-F1) por época sobre el corpus fusionado (6,903 tweets, 76/24, class weights). La línea punteada indica el umbral Macro-F1 ≥ 85%.

5.3. BETO

Se aplicó ajuste fino del modelo dccuchile/bert-base-spanish-wwm-cased [9] (110M parámetros), con lr=1e-5, AdamW con weight decay=0.05, dropout=0.2, warmup del 20%, tamaño de lote de 16 y máximo de 3 épocas con early stopping sobre val_loss (paciencia=2).

5.4. XLM- RoBERTa large

Se utilizó el modelo xlm-roberta-large [11] (560M parámetros, entrenado en 100 idiomas), con una configuración idéntica a la de BETO para garantizar una comparación controlada.

Todos los experimentos se realizaron en Google Colaboratory con GPU NVIDIA Tesla T4 (16 GB), Python 3.12, TensorFlow 2.19/Keras para LSTM y CNN, y PyTorch 2.x con Transformers 4.x para los modelos Transformer.

Por restricciones computacionales — especialmente el costo de XLM-RoBERTa large (~40 min/época en T4) — cada arquitectura se evaluó con su configuración óptima en una partición fija. La aplicación de validación cruzada k-fold y múltiples semillas se propone como trabajo futuro para habilitar pruebas de significancia estadística.

6. Resultados y discusión

6.1. Curvas de entrenamiento — BETO

La Fig. 1 muestra las curvas de entrenamiento de BETO sobre el corpus fusionado. La pérdida de validación decrece consistentemente durante las tres épocas sin divergencia respecto a la pérdida de entrenamiento, evidenciando ausencia de sobreajuste.

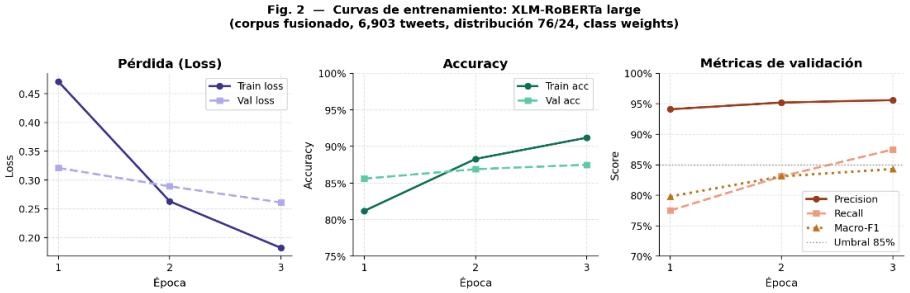

Fig. 2. Curvas de entrenamiento de XLM-RoBERTa large. Pérdida, Accuracy y métricas de validación por época sobre el corpus fusionado. La línea punteada indica el umbral Macro-F1 $\geq$ 85%.

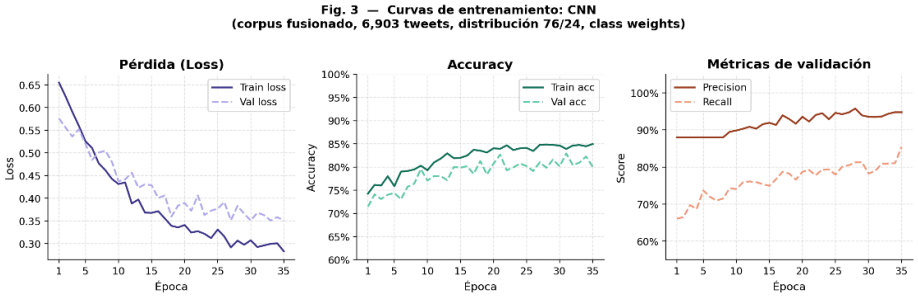

Fig. 3. Curvas de entrenamiento de la CNN. Pérdida (Loss), Accuracy, Precision y Recall por época sobre el corpus fusionado (6,903 tweets, 76/24).

El Macro-F1 en validación supera el umbral del 85% desde la época 2 (0.861), estabilizándose en 0.871 en la época 3. La convergencia estable y rápida se atribuye al preentrenamiento en 3,000 millones de palabras en español, que proporciona representaciones contextuales ricas desde el inicio del fine-tuning.

6.2. Curvas de entrenamiento — XLM-RoBERTa large

La Fig. 2 presenta las curvas de entrenamiento de XLM-RoBERTa large. Al igual que BETO, las curvas muestran convergencia estable sin sobreajuste. Sin embargo, la Precision se mantiene cercana a 0.96 durante las tres épocas mientras el Recall parte de 0.775 y escala hasta 0.875, sugiriendo que el modelo tarda más en aprender representaciones discriminativas para la clase minoritaria.

El Macro-F1 en validación llega a 0.843 en la época 3, sin alcanzar el umbral del 85%. Este comportamiento refleja las limitaciones del preentrenamiento multilingüe para capturar el matiz lingüístico específico del español mexicano con solo 3 épocas de fine-tuning.

Tabla 2. Resultados en conjunto de prueba — dataset fusionado con class weights.

Modelo	Acc. (%)	Macro-F1 (%)	Prec. (%)	Rec. (%)	FP	FN	Hipótesis
BETO	88.42	85.28	95.13	89.21	29	85	Sí (>85%)
XLM-RoBERTa large	87.45	84.13	95.57	87.56	32	98	No
CNN	81.85	77.99	93.99	81.35	41	147	No
LSTM	—	—	—	—	—	—	No converge

6.3. Curvas de entrenamiento — CNN

La Fig. 3 muestra las curvas de entrenamiento de la CNN a lo largo de 35 épocas. Las curvas de pérdida de entrenamiento y validación convergen de manera alineada desde las primeras épocas, confirmando la ausencia de sobreajuste tras la eliminación de la regularización L2.

La Precision de validación se estabiliza en torno a 0.950 desde la época 10, mientras que el Recall converge a 0.810, reflejando la tendencia de la CNN a priorizar la clase mayoritaria (misógino) al capturar patrones léxicos recurrentes vía filtros convolucionales de n-gramas.

La convergencia en ~35 épocas es notablemente más lenta que los modelos Transformer (3 épocas), aunque el costo computacional por época es significativamente menor.

7. Comparación de resultados

La Tabla 2 resume los resultados de las cuatro arquitecturas sobre el conjunto de prueba (1,036 mensajes).

Para LSTM se reportan los valores cuantitativos observados en todas las configuraciones exploradas: en todos los casos el modelo presentó TN≈0 (falsos negativos tendentes a cero para la clase positiva), Recall de clase minoritaria≈0.00 y Macro-F1<50%, equivalente a clasificación aleatoria sesgada hacia la clase mayoritaria.

BETO es el único modelo que confirma la hipótesis (Macro-F1 ≥ 85%), alcanzando 85.28% con solo 29 falsos positivos y 85 falsos negativos. XLM-RoBERTa large obtiene 84.13% de Macro-F1, inferior en 1.15 puntos a pesar de sus 560M parámetros. Este resultado es consistente con la literatura que documenta la superioridad de modelos monolingüe sobre multilingüe para tareas de PLN en lenguas específicas [12]: la especialización de BETO en el corpus español — incluyendo lenguaje coloquial y argot

de Twitter — proporciona representaciones semánticas más ajustadas que el preentrenamiento multilingüe generalista. Cabe destacar que XLM-RoBERTa obtiene la mayor Precision (95.57%, 32 FP), posicionándose como alternativa preferible cuando minimizar la censura de contenido legítimo es prioritario.

CNN alcanza Macro-F1 de 77.99% con ~1.4M parámetros, representando el 91.5% del Macro-F1 de BETO con el 1.3% de sus parámetros, confirmando su posición como mejor alternativa en términos de balance costo-rendimiento. El modelo LSTM con embeddings aleatorios (sin preentrenamiento) actúa implícitamente como línea base sin preentrenamiento, ilustrando el beneficio absoluto del preentrenamiento: frente al colapso total de LSTM, CNN —con embeddings entrenables desde cero— logra 77.99% de Macro-F1, y los modelos Transformer con preentrenamiento masivo en español alcanzan 84–85%.

Los falsos negativos de BETO (85 casos) corresponden predominantemente a tweets donde el lenguaje misógino aparece codificado en ironía, humor o argot coloquial mexicano no recurrente en el corpus de preentrenamiento. Los falsos positivos (29 casos) corresponden mayoritariamente a tweets que discuten o citan violencia de género usando vocabulario misógino sin ser ellos mismos misóginos. Por restricciones de privacidad y extensión, no se reproducen ejemplos directos de tweets, pero estos patrones de error sugieren que técnicas de augmentación de datos orientadas a ironía y lenguaje coloquial podrían mejorar los resultados futuros.

8. Conclusiones

Este trabajo presentó una comparación sistemática de cuatro arquitecturas de redes neuronales unimodales para la detección automática de mensajes misóginos en español mexicano, abordando el problema del desbalance de clases emergente al fusionar los datasets MEX-A3T y EXIST 2024. Los resultados conducen a cuatro conclusiones principales.

Primero, la especialización lingüística en español es más determinante que la capacidad paramétrica: BETO (110M parámetros) supera a XLM-RoBERTa large (560M parámetros) en Macro-F1 (85.28% vs 84.13%) y en reducción de errores (29 vs 32 FP; 85 vs 98 FN). Este hallazgo es consistente con evidencia previa en la literatura [12] y aporta sustento empírico específico para el dominio de detección de misoginia en español mexicano.

Segundo, los pesos de clase proporcionales superan al undersampling preservando el volumen de información de entrenamiento y obteniendo mejoras de hasta 1.56 puntos en Macro-F1 respecto a la reducción del corpus. Tercero, LSTM sin preentrenamiento en español no logra convergencia estable ante desbalance 1:3.18 con corpus moderado, delimitando las condiciones bajo las que las arquitecturas recurrentes resultan viables para detección de discurso de odio. Cuarto, CNN con ~1.4M parámetros representa la alternativa con mejor balance costo-rendimiento, alcanzando el 91.5% del Macro-F1 de BETO con el 1.3% de sus parámetros.

Como trabajo futuro se propone explorar: (a) modelos RoBERTa monolingüe en español como MarIA-BNE; (b) data augmentation textual para la clase minoritaria; (c)

clasificación multiclase de categorías de misoginia; y (d) evaluación en corpus de otras plataformas.

Referencias

1. Fersini, E., Rosso, P., Anzovino, M.: Overview of the Task on Automatic Misogyny Identification at IberEval 2018. IberEval@SEPLN 2150, pp. 214–228 (2018)
2. WHO: Global and Regional Estimates of Violence against Women. World Health Organization, Geneva (2013)
3. Molina-Villegas, A.: La incidencia de las voces misóginas sobre el espacio digital en México. In: Jóvenes, Plataformas Digitales y Lenguajes. Elementum, Pachuca (2021)
4. Hewitt, S., Tiropanis, T., Bokhove, C.: The problem of identifying misogynist language on Twitter. In: ACM Web Science, pp. 333–335 (2016)
5. Vera Lagos, V.: Detección de misoginia en textos cortos mediante clasificadores supervisados. Tesis de Licenciatura, BUAP, México (2021)
6. Frenda, S., Bilal, G.: Exploration of Misogyny in Spanish and English Tweets. In: IberEval 2018, Vol. 2150, pp. 260–267 (2018)
7. García-Díaz, J.A., Cánovas-García, M., Colomo-Palacios, R., Valencia-García, R.: Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. Future Gener. Comput. Syst., Vol. 114, pp. 506–518 (2021)
8. Aldana-Bobadilla, E., Molina-Villegas, A., Montelongo-Padilla, Y., Lopez-Arevalo, I., Sordia, O.S.: A Language Model for Misogyny Detection in Latin American Spanish Driven by Multisource Feature Extraction and Transformers. Appl. Sci., Vol. 11, 10467 (2021)
9. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: PML4DC at ICLR (2020)
10. Rodríguez, D.A., Diaz-Escobar, J., Díaz-Ramírez, A. et al.: Domain-adaptive pre-training on a BERT model for the automatic detection of misogynistic tweets in Spanish. Soc. Netw. Anal. Min., Vol. 13, pp. 126 (2023)
11. Conneau, A., et al.: Unsupervised Cross-lingual Representation Learning at Scale. In: ACL 2020, pp. 8440–8451 (2020)
12. Salmerón-Ríos, S., et al.: Evaluation of transformer models for tasks in Spanish. PeerJ Comput. Sci. (2024). Doi:10.7717/peerj-cs.1992.
13. Plaza-del-Arco, F.M., et al.: Overview of EXIST 2024 — Learning with Disagreement for Sexism Identification and Characterization. In: CLEF (2024)
14. Kim, Y.: Convolutional Neural Networks for Sentence Classification. In: EMNLP, pp. 1746–1751 (2014)
15. Pereira, A.F., et al.: AI-UPV at IberLEF-2021 EXIST task: Sexism Prediction Using Monolingual and Multilingual BERT. In: IberLEF@SEPLN (2021)