VISUAL REVIEW | Vol. 18, No. 1, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.6128

GESTIÓN DE RECURSOS HUMANOS EN LA MODERACIÓN DE CONTENIDOS DE REDES SOCIALES
Perspectivas desde la Ley de Servicios Digitales

MARIAN NÚÑEZ-CANSADO[1]
mariangeles.nunez@uva.es
ÁNGEL QUINTANA GÓMEZ [2]
angel.quintana@ulpgc.es
[1] Universidad de Valladolid, España
[2] Universidad de las Palmas de Gran Canarias, España

PALABRAS CLAVE

Moderación de contenidos
Redes Sociales
Ley de servicios digitales
Ética de la comunicación
Unión Europea
Comunicación
Publicidad

RESUMEN

Este estudio evalúa la implementación de recursos humanos en la moderación de contenidos en grandes plataformas digitales (VLOPs) dentro de la Unión Europea, utilizando una metodología cuantitativa basada en análisis documental y de datos. Los resultados muestran disparidades significativas en la ratio de moderadores por idioma, con TikTok y YouTube liderando en cobertura, mientras que LinkedIn y X presentan vacíos críticos. Se concluye que, aunque las plataformas cumplen con los dictámenes regulatorios, la falta de cobertura lingüística completa y las condiciones laborales inadecuadas representan riesgos significativos para comunidades vulnerables y moderadores.

Received: 13/ 07 / 2025
Accepted: 28/ 10 / 2025

1. Introducción

La capacidad de almacenar y compartir contenidos en línea a gran escala ha transformado significativamente la forma en que las personas se relacionan entre sí, otorgándoles un papel crucial en la comunicación personal, la construcción de identidad y la expresión cultural (Manovich, 2020; Mitchell, 2017). Sin embargo, este raudal de contenidos presenta importantes dilemas éticos y de privacidad, especialmente en lo que respecta a la gestión de imágenes digitales, en aspectos como la selección, organización y exhibición de imágenes así como por su posible naturaleza violenta y/o de incitación al odio.

La edición de contenidos en medios digitales genera dilemas relacionados con los derechos de autor y el consentimiento de las personas fotografiadas, así como con el manejo de datos personales que podrían estar involucrados en los contenidos utilizados, entre otros. Es importante destacar que aunque las redes sociales permiten a los usuarios compartir contenido de manera rápida y masiva, estas plataformas incluyen en sus términos y condiciones cláusulas que otorgan a la red social una licencia de propiedad amplia sobre el contenido publicado por los usuarios. Este mecanismo contractual, tal y como afirma Hernández García (2021), implica que los usuarios ceden derechos de uso a la plataforma, lo que permite su reutilización dentro de los límites definidos en el acuerdo de uso. Existe numerosa literatura científica que destaca la problemática del uso no autorizado de contenidos y la vulneración de la privacidad individual, que a menudo puede comprometer la seguridad del usuario de las redes (Nissenbaum, 2011). Junto a estos problemas, debemos resaltar que la reutilización y la falta de contexto en la divulgación en plataformas digitales pueden dar lugar a problemas de plagio y distorsión de la representación, lo que puede afectar negativamente la percepción pública de ciertos individuos y/o grupos sociales (Boyd & Marwick, 2017). Por último no debemos dejar de lado aquellos contenidos que promueven la violencia, la incitación al odio y la explotación infantil, lo cuales pueden generar un impacto profundamente perjudicial tanto en la construcción de identidad individual del sujeto como en la de la identidad colectiva.

Es importante reconocer que esta situación es especialmente sensible en el caso de los niños, que conforman un grupo particularmente vulnerable a las implicaciones de la difusión de contenidos digital (Lievens et al., 2019; Livingstone & Third, 2017). Algunos de estos riesgos, por ejemplo, que pueden hacerse patentes en la explotación digital de la imagen e identidad, que pueden acarrear la posibilidad de desarrollar un autoconcepto distorsionado, que genere serios problemas psicosociales (Núñez-Cansado et al., 2021).

A los múltiples riesgos asociados con los contenidos presentes en las redes sociales y el derecho de privacidad, es fundamental añadir el impacto de los algoritmos utilizados por las plataformas. Estos sistemas, diseñados para maximizar la interacción y el tiempo de permanencia de los usuarios, suelen priorizar contenidos sensacionalistas, polarizantes o emocionalmente impactantes. En el año 2021, un exempleado de la corporación Facebook-Meta puso en entredicho a su directiva al filtrar decenas de documentos al diario *The Wall Street Journal*. Estos documentos revelaban prácticas cuestionables que favorecían una influencia negativa sobre la población más joven, derivada de los contenidos y algoritmos implementados por la plataforma. Tal y como declaró el trabajador en su testimonio ante el Parlamento Europeo:

"Los modelos algorítmicos empleados por redes sociales como Instagram estaban especialmente diseñados para favorecer artificialmente la comparación social, erosionando la autopercepción de mujeres jóvenes sobre sus cuerpos, prácticas sociales y recursos económicos" (Jiménez González & Cancela Rodríguez, 2023, p. 91).

Arturo Béjar, líder del departamento *'Integrity and Care*' de Facebook, revelaba en el documental; Redes Sociales : la fábrica del terror, que la plataforma conocía e ignoraba los daños a sus usuarios. El ex directivo, envió, en junio de 2020, el renombrado mail a Mark Zuckerberg y al resto de los altos cargos de la corporación en el que aportaba datos escalofriantes, como que el 51% de los usuarios de Instagram reportaban contenidos negativos o que el 21,8% de adolescentes de 13 a 15 años reportaban haber recibido acoso directo, afirmando tajantemente que probablemente Instagram fuera el responsable del mayor acoso sexual de la historia de la humanidad (Évole Requena & Lara, 2024).

Se destapa así una realidad que impulsa un nuevo debate sobre los peligros en las redes y que han generado la necesidad de instrumentar sistemas que regulen los contenidos difundidos por las plataformas.

Aunque existían previamente disposiciones jurídicas sobre el uso y la protección de datos en internet - en particular el Reglamento General de Protección de Datos (GDPR) de Europa- estos reglamentos se presentaban como insuficientes ante el perjuicio denunciado. Así, Alemania promulgó en 2017 La Ley de Redes, más conocida como NetzDG, que fue la base para diferentes disposiciones como el Código Europeo de Buenas Prácticas para la Desinformación, aprobado por la Unión Europea en el año 2018, o la Disposición ministerial aprobada por el Consejo de Seguridad Nacional de España el 30 de octubre de 2020, en la que se establece el Procedimiento de Actuación contra la Desinformación. La Unesco, se sumó a la preocupación y desarrolló un Seminario sobre la Moderación de contenidos y libertad de expresión en las redes sociales, en el que se abordó la regulación de los contenidos y la protección de los derechos humanos, indagando en temas como los procedimientos: IA VS recursos humanos, los criterios de moderación, la transparencia y la necesidad de la rendición de cuentas por parte de las grandes plataformas a las autoridades pertinentes.

La ausencia de una regulación específica y el desafío que suponía la necesidad de moderar los contenidos digitales con el fin de mitigar los riesgos asociados, forzó el compromiso de corporaciones y gobiernos, que se materializó en la promulgación, por parte de la Unión Europea, de la Ley de Servicios Digitales (DSA) que entraría en vigor el 25 agosto de 2023 (Unión Europea, 2022). La DSA, se presenta como una pieza clave para generar reglas claras que garanticen un entorno digital seguro. Esta ley establece obligaciones específicas que deben respetar las corporaciones, tales como la transparencia en los algoritmos, la prohibición de prácticas publicitarias intrusivas engañosas, el fortalecimiento del mecanismo de denuncias, la eliminación de contenidos ilícitos e introduce además, medidas para proteger a los usuarios de contenidos perjudiciales, especialmente a los menores. Según la normativa las grandes plataformas (VLOPs) con más de 45 millones de usuarios activos dentro de la Unión Europea, deberán tomar medidas adicionales para gestionar los riesgos sistémicos derivados de su actividad, y estarán obligadas a rendir cuentas mediante informes periódicos sobre la moderación de contenidos y la publicidad política. El objetivo básico de esta ley es equilibrar la seguridad, la transparencia y la responsabilidad en el ecosistema digital de la Unión europea, responsabilizando a las plataformas en cuanto a los contenidos que albergan y protegiendo a los usuarios frente a contenidos dañinos, algoritmos perjudiciales y el control abusivo de los datos personales.

Sin embargo, estas leyes aún enfrentan importantes desafíos en su implementación y aplicación dentro de un contexto de transformación digital constante, donde la moderación de contenidos suele resultar compleja, ya sea por conflictos con los intereses comerciales o por la ausencia de regulaciones específicas en determinados espacios virtuales (Jiménez González & Cancela Rodríguez, 2023).

2. Moderadores de contenido

Los riesgos que pueden acarrear el mal uso del contenido en las redes sociales han incrementado la necesidad, no solo de un análisis profundo de la legislación, sino también del compromiso por parte de las corporaciones de medios sociales. La variedad de contenidos, los algoritmos de recomendación y gráficos sociales han dado lugar a múltiples quejas y críticas hacia las plataformas de contenidos, circunstancia que ha obligado a las corporaciones de medios a mantener un compromiso con sus usuarios de monitoreo constante y riguroso, lo que plantea desafíos tanto logísticos como de relaciones públicas (Gillespie, 2020). Es así como cobra vital importancia la “moderación de contenido comercial” o “moderación de plataformas”. En los últimos años se ha producido un gran incremento de literatura científica que aborda los principales problemas sobre este nuevo perfil profesional que va desde la preocupación por la situación laboral, el análisis de la legislación, la legitimidad democrática e incluso la falta de transparencia y responsabilidad de las grandes corporaciones (Gorwa et al., 2020).

La moderación de contenidos es la práctica organizada del filtrado de contenidos generados por los usuarios que publican en sitios de internet (Roberts, 2019), es definida como: “mecanismos de gobernanza que estructuran la participación en una comunidad para facilitar la cooperación y prevenir el abuso" (Grimmelmann, 2017: 18). Su objetivo es garantizar un entorno seguro y respetuoso para todos los usuarios de las redes, evitando la difusión de contenido inapropiado, ilegal o perjudicial, como discursos de odio violencia, desinformación, acoso o material explícito. El material puede ser revisado previa distribución, o puede ser evaluado como resultado de una queja de un usuario, de un administrador del sitio, de empresas afectadas…

Este nuevo perfil profesional ha acarreado en el último año debates sobre tres aspectos fundamentales, la creación de manuales básicos que permitan el desarrollo de su profesión, el reconocimiento de los riesgos psicosociales, y consecuentemente el reconocimiento de los efectos como enfermedades laborales, y las normas de convivencia de regulación humana Vs regulación IA.

La Agencia Europea para la Seguridad y Salud en el trabajo (OSHA) publicó en el año 2022 el documento: *Occupational safety and health risks of online content review work provided through digital labour platforms*, en el que se manifestaba como fuente de estrés la exposición a contenidos violentos, crimen, abuso y contenidos ilegales, que podrían originar daños psicológicos y trastornos postraumáticos (Lousada, 2024), avalando la decisión del Juzgado de lo Social núm. 28 de Barcelona emitida en el año 2024 sobre la clasificación de incapacidad temporal de un moderador de contenidos que desempeñaba su trabajo en la empresa CCC Barcelona Digital Services. El trabajador presentaba ataques de pánico, conductas de evitación, aislamiento en el domicilio, rumiaciones hipocondriacas, sensación de disfagia, despertares nocturno e importantes tanatofobias. En la propia página de la empresa se difundía un contenido que afirmaban conocer las consecuencias del ejercicio de la profesión y anunciaban su prioridad por defender la salud mental y el bienestar de sus trabajadores por lo que ponían a la disposición de su empleados tecnología e implementaban moderación con inteligencia artificial (Lousada Arochena, 2024). En el año 2022 La plataforma Meta, en un reconocimiento histórico acordó pagar 52 millones de dólares a moderadores tanto actuales como anteriores para compensar los problemas de salud mental desarrollados en su trabajo (Newton, 2020).

La necesidad de paliar los efectos sobre la salud mental de los trabajadores ha impulsado la creación de algoritmos que facilitan, a priori, la función de revisar y clasificar el material en las redes, a partir de la utilización de los datos de participación de los usuarios (Wang et al., 2022) o de la creación de redes neurales y sistemas de Inteligencia artificial (Agarwal et al., 2020; Alharthi et al., 2021; Andročec, 2020). Estos nuevos mecanismos de moderación prometían la panacea y la solución a la grave crisis de descrédito que estaba sufriendo el sector. Mark Zuckerberg, resaltó docenas de veces durante su testimonio en el congreso en 2018, el papel de la IA como futura solución a la moderación de contenidos (Gorwa et al., 2020).

En marzo de 2020, ante el confinamiento de la población en respuesta a la crisis sociosanitaria de la COVID, Twitter y otras muchas corporaciones se vieron obligadas, a utilizar exclusivamente la moderación automatizada. El sistema mostró unas graves deficiencias y tras el fracaso, muchas plataformas se disculparon por el aumento de errores (Gillespie, 2020). El debate sobre la moderación de contenidos mediante la IA se abre con detractores y defensores, la Declaración de la IFLA sobre Bibliotecas e Inteligencia artificial (Committee on Freedom of Access to Information, 2020) concluye sobre esta cuestión que el uso de IA en la moderación de contenidos supone una infracción contra la libertad de expresión, que puede conducir a la falta de transparencia y opacidad con relación a los criterios de eliminación. Uno de los casos más representativos de este riesgo fue sacado a la luz en el año 2020 mediante la filtración de las pautas de moderación de contenidos marcadas por la plataforma Tik Tok, donde se instaba a la IA a eliminar las publicaciones creadas por usuarios "poco Atractivos" físicamente, lo que incluía a personas con discapacidad (Torres Vargas, 2022).

La necesidad de establecer normas claras, que guíen tanto a la regulación de los contenidos mediante algoritmos de IA como a los trabajadores humanos, ha llevado a las plataformas a publicar políticas de directrices que podemos resumir y aunar en los siguientes puntos:

Contenido de alta prioridad: explotación y abuso sexual infantil, violencia gráfica, discursos de odio, acoso o *bullying*, contenido extremista o terrorista, contenido relacionado con el suicidio o la autolesión, desinformación peligrosa, contenido que incitan a la violencia o el crimen, contenido relacionado con drogas, discurso de odio.

Contenidos de prioridad moderada: Desinformación, contenido relacionado con drogas ilegales, abuso de animales, violencia basada en género, lenguaje ofensivo, suplantación de identidad, contenido gráfico perturbador.

Contenidos de prioridad baja: spam y fraude, Teorías de conspiración, contenido de broma o sátira dañinas, violencia baja o situaciones perturbadoras.

Los trabajadores pueden ser asignados a una de estas modalidades de contenidos o incluso estar asignados a todas ellas, es complejo acceder a esta información dado los contratos de confidencialidad que han de firmar. Esta confidencialidad es uno de los grandes problemas pues existe escasa información sobre sus condiciones de trabajo, dinámicas laborales y en general políticas que utilizan las

grandes plataformas. Estas cláusulas permiten a las organizaciones mantener una praxis con una gran falta de transparencia (Torres Vargas, 2022)

Como una alternativa a la moderación de contenidos automatizada y para fomentar un enfoque más inclusivo, X amplió su iniciativa conocida como Birdwatch, transformándola en Community Notes. Este sistema permite a la comunidad de usuarios evaluar y contextualizar la calidad del contenido difundido, ya sea texto, imagen o video, mediante notas colaborativas visibles para otros usuarios en la plataforma (Chuai et al., 2024). Esta acción se implementa inicialmente en Estados Unidos en 2023 y llega a España en 2024, sin incumplir las medidas establecidas por la DSA, ya que continúa presentando sus informes de transparencia, en los cuales se indica: "Nuestros sistemas de moderación de contenidos combinan la revisión automatizada y humana con un sólido sistema de apelación que permite a nuestros usuarios plantear rápidamente posibles anomalías o errores de moderación" (X Corp, 2024, s.p.).

En 2025, algunas organizaciones dan un paso más en la moderación. En enero, el CEO de Meta, Mark Zuckerberg, realizó un comunicado a través de Facebook en el que anunció que pone fin a los moderadores de contenidos, sustituyéndolos por un sistema de moderación basado en los propios usuarios de la plataforma, siguiendo un modelo similar al implementado por X en Estados Unidos. Sin embargo, continuará moderando contenido relacionado con drogas, terrorismo y explotación infantil (Zuckerberg, 2025). Alexios Mantzarlis, director de Seguridad, Confianza y Protección en la Universidad de Cornell, citado en *El País* por Limón (2025) señala que "la decisión de no verificar datos y de relajar la moderación de contenidos abre la puerta a un incremento del discurso de odio".

Los múltiples daños que puede acarrear la difusión de contenido perjudicial, y la falta de seguridad en las redes, en un momento tan cambiante y de gran incertidumbre respecto a las decisiones de plataformas y organismos públicos, hacen necesario un análisis de la gestión de recursos humanos en la moderación de contenidos desde la perspectiva de la Ley de Servicios Digitales.

3. Objetivos

Evaluar la implementación de recursos humanos en la moderación de contenidos en las grandes plataformas digitales (VLOPs) con más de 45 millones de usuarios, desde la perspectiva de la Ley de Servicios Digitales.

Este dato nos permitirá conocer si las redes sociales asignan una cantidad adecuada de moderadores en relación con sus usuarios activos/idiomas en la Unión Europea (UE)

4. Metodología

Se emplea una metodología cuantitativa basada en un enfoque de análisis documental, revisión de informes, análisis de datos y evaluación de los recursos humanos utilizados en la moderación contenidos de las plataformas VLOPs, con funcionamiento activo en la Unión Europea.

La investigación se lleva a cabo en dos fases:

1º Fase : Revisión de informes de transparencias del segundo cuatrimestre del año 2023 de las principales plataformas digitales (VLOPs) activas en la Unión Europea. Se trata de analizar los recursos humanos empleados en la moderación de contenido a partir del primer informe resultante de la Ley de Servicios Digitales (Comisión europea, 2023)

Fuente de datos: Primer informe de transparencia, en cumplimiento de la DSA, correspondiente al segundo cuatrimestre del 2023. Se utilizan los informes de las principales plataformas: YouTube, Meta (Facebook, Instagram) Pinterest, LinkedIn, Snapchat y X.

El objetivo es extraer información cuantitativa sobre la gestión de recursos humanos en la plataforma, prestando especial atención a la cobertura de idiomas oficiales de la organización europea.

2º Fase: Análisis y síntesis de los datos obtenidos de los informes. Llevamos a cabo estadísticas descriptivas, medidas de proporción, índices de equilibrio. El objetivo es integrar los resultados cuantitativos obtenidos en las fases anteriores para desarrollar una comprensión integral sobre el uso de recursos humanos en la moderación de contenidos en plataformas digitales.

5. Análisis de Datos

A los datos cuantitativos extraídos de los informes de las grandes plataformas atendiendo al número de moderadores por idiomas, se le ha aplicado la evaluación de la razón de moderadores/usuarios

llevada a cabo mediante el cálculo de la ratio de la proporción de moderadores por usuario de cada red social utilizando la siguiente fórmula:

R mu = \frac{M}{U}

Esta ratio (R mu) genera una métrica estandarizada que facilita la comparación entre las diferentes plataformas. Los resultados se muestran en la tabla 1:

Tabla 1. Razón moderadores por usuarios (Rmu)

Red Social	R mu[1]
X	11.81
META	4.83
TIK TOK	41.85
YouTube	4.15
Snapchat	17.49
Pinterest	0.97
LinkedIn	3.25

Fuente: elaboración propia, 2024.

Para evaluar la comparación relativa (Rmu), se evalúa la consistencia en las proporciones R de las plataformas, resultado que arroja diferencias notables. El promedio general se sitúa en 10.27 moderadores por millón de usuarios destacando una variabilidad extremadamente alta en la dispersión, lo que indica una distribución no homogénea ($\sigma = 322289.72$), reflejando una gran disparidad en la asignación de moderadores por usuarios con algunas plataformas excediendo significativamente el promedio (CV=31347.49%).

La razón por idiomas nos muestra resultados muy dispares. (Tabla 2), el idioma inglés lidera el promedio ($\bar{x}$=298.15) mucho más que cualquier otro idioma y con un rango extremo (998.6), debido sobre todo a valores extremados en Tik Tok y Snapchat. Los idiomas mejor balanceados son el español y el portugués con promedios altos y medianas cercanas, que muestran una cobertura más uniforme. Idiomas como el lituano y el griego tienen bajos promedios y medianas nulas, indicando poca inversión en recursos humanos.

Tabla 2. Promedios, mediana y rangos por idiomas

Idioma	**Promedio de *R***	**Mediana de *R***	**Rango (*max−min*)**
Alemán	8.21	3.62	37.60
Búlgaro	6.28	1.42	36.32
Checo	5.53	0.00	32.63
Croata	5.58	0.00	22.22
Danés	6.01	1.80	32.31
Esloveno	15.12	0.00	90.00
Eslovaco	7.15	0.00	48.89
Estonio	3.34	0.00	15.00
Español	8.98	6.33	24.64
Finlandés	4.93	2.83	23.67
Francés	6.73	2.61	27.23
Griego	3.12	0.00	16.27
Holandés	4.55	0.81	23.86
Húngaro	4.27	0.00	23.33
Inglés	298.15	25.35	998.60
Italiano	5.07	2.50	22.16
Letón	10.97	1.00	61.11
Lituano	2.02	0.00	7.50
Polaco	4.44	2.77	19.62
Portugués	16.44	13.33	45.05
Rumano	4.21	1.81	22.57
Sueco	6.19	1.50	33.75

Fuente: elaboración propia, 2024.

Para analizar la distribución de la razón entre los diferentes idiomas de la Unión Europea (Rmui), se calcula la ratio de la proporción de moderadores por usuarios según idiomas (tabla 3).

[1] Ratio Moderadores por millón de usuarios

Según esta razón, el idioma inglés destaca por encima del resto (Rmui=298.15) posiblemente debido a los valores extremadamente altos en la red Tik Tok. El portugués (Rmui= 16.44) y el esloveno (Rmui=15.12) muestran valores con una moderación media. El lituano (Rmui= 2.02) el griego (Rmui=3.12) y el estonio (Rmui=3.34) se encuentra entre los idiomas con las ratios más bajas.

Las redes Tik ToK ($\sigma = 207.5$) y Snapchat ($\sigma = 124.8$) muestran una gran variabilidad en los valores (Rmui) entre idiomas, esto indica que aunque mantiene buenos promedios su cobertura no es uniforme y existen idiomas con valores muy altos y muy bajos. La red X también mantiene una alta variabilidad ($\sigma = 97.06$), presentan valores muy elevados en el idioma inglés, mientras que mantiene valores 0 en una gran parte de los idiomas.

META ($\sigma = 5.04$) , Pinterest ($\sigma = 4.57$) y LinkedIn ($\sigma = 5.50$) muestran menos dispersión en la ratio, lo que indica una cobertura más uniforme, aunque posiblemente limitada en algunos idiomas.

La desviación más sobresaliente se produce en Tik Tok, pues presenta una desviación estándar muy elevada que refleja un enfoque intensivo en algunos idiomas pero falta de cobertura en otros.

Tabla 3. Razón moderadores por usuarios (Rmui)/idiomas

Red	R_{mu} /redes[2]						
	X	META	TIK TOK	YouTube	Snapchat	Pinterest	LinkedIn
Alemán	3.76	6.47	37.78	2.54	3.62	0.18	3.12
Búlgaro	1.42	4.55	36.32	1.67	0	0	0
Checo	0	2.97	32.63	3.13	0	0	0
Croata	0	11.36	22.22	5.45	0	0	0
Danés	0	3.95	32.31	1.8	4.02	0	0
Esloveno	0	7.5	90	15	8.33	0	0
Eslovaco	0	0	48.89	1.16	0	0	0
Estonio	0	3	15	5.38	0	0	0
Español	0.93	4.72	25.57	10.02	13.92	1.40	6.33
Finlandés	0	4.69	23.67	2.83	3.35	0	0
Francés	2.61	4.12	27.59	2.13	7.68	0.36	2.61
Griego	0	2.97	16.27	2.59	0	0	0
Holandés	0	5.05	23.86	0.81	0	0.2	1.9
Húngaro	0	3.48	23.33	3.09	0	0	0
Inglés	46	25.95	101.7	2.46	587.20	18.27	0
Italiano	0.12	5.03	22.28	1.68	3.69	0.2	2.5
Letón	0	1.82	12.86	61.11	0	1.1	0
Lituano	0	2.86	7.5	3.79	0	0	0
Polaco	0	2.77	19.62	2.99	0.69	0	5
Portugués	2.02	6.58	22.73	45.05	0	13	25
Rumano	0	2.87	22.57	2.21	1.81	0	0
Sueco	0	2.96	33.75	1.5	4.56	0.5	0

Fuente: Elaboración propia, 2024.

La comparación de cobertura por red social arroja datos muy relevantes (tabla 4), las redes Tik Tok y YouTube presenta una cobertura prácticamente completa, ambas redes cubren 22 idiomas con al menos un moderador. Meta destaca con un 95.45% de cobertura, tan solo falta la cobertura en un idioma de los 24 oficiales en la Unión europea, teniendo en cuenta que el maltés no es contemplado en ningún caso y el irlandés forma se asume dentro de la lengua inglesa. El caso de Snapchat (45.45%), Pinterest (40.91%) y LinkedIn (31.82%), las coberturas son más limitadas, carecen de moderadores en la mitad de los idiomas. Por último la red X (31.82%) presenta una baja cobertura, deja sin moderadores una gran parte de idiomas.

Tabla 4. Comparación de cobertura por red social

Red Social	Idiomas con Cobertura (R>0)	Idiomas sin Cobertura (R=0)	Porcentaje de Cobertura (%)
X	7	15	31.82%
META	21	1	95.45%
Tik ToK	22	0	100.00%
YouTube	22	0	100.00%

[2] 0: plataformas con 0 moderadores

Snapchat	10	12	45.45%
Pinterest	9	13	40.91%
LinkedIn	7	15	31.82%

Fuente: elaboración propia, 2024.

El rendimiento relativo de los moderadores nos permite tener una idea de la distribución de los recursos de moderadores según redes y usuarios (tabla 5).

Así Tik Tok tiene el mayor rendimiento relativo (72.52) combinando una alta cobertura lingüística con una gran inversión en moderadores, lo que sugiere un enfoque bien distribuido. La red X (66.66) y Snapchat (63.05) presenta también altos valores, aunque con menor cobertura lingüística. Lo que indica un enfoque más limitado. Los valores más bajos los presentan Meta (5.48) YouTube (7.81), Pinterest (3.88) y LinkedIn (6.69), con una distribución más uniforme, pero con menos intensidad por idioma.

Tabla 5. Rendimiento relativo de moderadores

Red Social	Total Moderadores (ΣR)	Rendimiento Relativo (Moderadores por Idioma con R>0)
X	466.59	66.66
META	115.07	5.48
Tik ToK	1595.35	72.52
YouTube	171.72	7.81
Snapchat	630.54	63.05
Pinterest	34.92	3.88
LinkedIn	46.85	6.69

Fuente: elaboración propia, 2024.

Esta variabilidad nos muestra que es posible que existan idiomas que dependan significativamente de una sola red, para verificar este dato se ha calculado el porcentaje total de cada idioma que corresponde a cada red social, este dato nos permitirá conocer la dependencia (Tabla 6). Según este índice la red social Tik Tok es la que más dependencia muestra para la mayoría de los idiomas, con una dependencia superior al 50%. YouTube destaca en letón (79.68%) y portugués (39.14))

Idiomas como el eslovaco y el esloveno dependen casi exclusivamente de Tik ToK, lo que podría suponer una limitación para estos países.

Tabla 6. Índice de dependencia por Idioma

Idioma	**Red con Mayor Dependencia**	**Porcentaje de Dependencia (%)**
Alemán	Tik ToK	65.74%
Búlgaro	Tik ToK	82.62%
Checo	Tik ToK	84.25%
Croata	Tik ToK	56.93%
Danés	Tik ToK	76.78%
Esloveno	Tik ToK	85.04%
Eslovaco	Tik ToK	97.68%
Estonio	Tik ToK	64.16%
Español	Tik ToK	40.66%
Finlandés	Tik ToK	68.53%
Francés	Tik ToK	58.58%
Griego	Tik ToK	74.53%
Holandés	Tik ToK	74.89%
Húngaro	Tik ToK	78.03%
Inglés	Tik ToK	47.85%
Italiano	Tik ToK	62.74%
Letón	YouTube	79.58%
Lituano	Tik ToK	53.00%
Polaco	Tik ToK	63.15%
Portugués	YouTube	39.14%
Rumano	Tik ToK	76.61%
Sueco	Tik ToK	77.94%

Fuente: elaboración propia, 2024.

6. Discusión

El análisis de los datos presentados nos arroja información sobre la desigualdad en las redes sociales con a la gestión de los recursos humanos empleados en la moderación de contenidos. Existen diferencias notables en la inversión de recursos asignados por las plataformas, lo que podría afectar a la equidad y la eficiencia de la moderación según lo exigido por la DSA. Tik Tok se mantiene como líder en promedio y en rendimiento relativo, a pesar de no tener una cobertura plena en los idiomas oficiales. Meta, YouTube y LinkedIn mantienen un nivel bajo de asignación de recursos. De todas las redes analizadas Pinterest mantienen una proporción que podría suponer una carencia en la asignación de recursos. Esta precariedad podría generar no solo problemas en la regulación de los contenidos, propiciando mensajes más violentos, o redes más inseguras, sino que además podría estar afectando a la carga de trabajo de los moderadores de contenido. En los últimos meses, son prolíferas las noticias con relación a los riesgos psicosociales que arrastran este perfil profesional, no solo debido a una sobrecarga laboral, sino también por la exposición constante a imágenes impactantes y emocionalmente agotadoras. Aún más si han de respetar los mínimos de rendimiento marcados por las propias corporaciones, situados en torno al 95%, de identificación correcta de contenido perjudicial (Dang et al., 2018).

Para poder analizar detenidamente la adecuación de recursos humanos en la moderación de contenidos sería indispensable conocer el volumen de producción de cada plataforma, pero este dato no consta en los informes de transparencia, tampoco supone un requisito en la Ley DSA. Si lo es, informar del número de usuarios a pesar de que conocer este dato puede ser poco útil a la hora de analizar los recursos, pues ha de distinguirse entre usuarios activo, usuarios pasivos y usuarios creadores de contenidos. Puede ocurrir que una red posea un elevado número de usuarios y que estos sean mayormente pasivos produciendo un volumen menor de contenidos que una red con menor número de usuarios pero con mayor actividad en la creación de contenidos. El valor del dato de la producción de cada plataforma debería ser tenido en cuenta en los informes de transparencia, no solo para certificar la seguridad en la red sino también para salvaguardar la salud psicosocial de los moderadores. Así por ejemplo, si tomamos los datos de volumen de videos publicados en YouTube (Ceci., 2023); 720.000 horas de contenido al día, y establecemos la ratio de horas que debiera evaluar cada moderador- sin tener en cuenta el idioma- tendríamos un total de 389.8 horas por moderados, razón totalmente fuera de los límites.

Otro dato de vital importancia para analizar la adecuada gestión de recursos sería examinar la distribución de moderadores vs idioma. Los resultados de este análisis reflejan un claro predominio de la moderación en el idioma inglés, que supera por amplio margen al resto, esto refleja una concentración significativa de recursos de moderadores en este idioma. Los idiomas como el portugués y el esloveno destacan con valores moderados lo que implica una cobertura aceptable en algunas redes, pero limitadas en otras. El lituano, el griego y el estonio tienen cifras significativamente bajas, e incluso presentando ausencia de moderadores en algunas redes.

Redes como Snapchat, Pinterest y LinkedIn, y X presentan una cobertura muy limitada, con menos de la mitad de idiomas representados, lo que sugiere una inversión insuficiente en moderación multilingüe.

Las redes que alcanzan mayor cobertura lingüística son Tik ToK y YouTube. Aunque en el caso de Tik Tok se produce una gran desigualdad en la intensidad de moderación según idioma, esta circunstancia podría deberse a una estrategia regional o comercial que prioriza ciertos mercados lingüísticos. Estos datos muestran además una dependencia extrema de la red Tik Tok ya que en los 20 de los 22 idiomas presentes, equivale a más del 50% del total de los moderadores asignados. Idiomas como el eslovaco y el esloveno, evidencian riesgos significativos en caso de cambio de política de moderación de esta red.

En el caso de YouTube destaca su papel en los idiomas letón y portugués, lo que podría significar una oportunidad para diversificar la cobertura en otros idiomas.

Con relación al equilibrio de cobertura por idioma, el español y portugués arrojan los valores más balanceados, con promedios altos y medianas cercanas, lo que refleja una moderación más uniforme en comparación con otros idiomas. Los promedios del lituano y el griego son los que presentan datos más preocupantes, pues muestra promedios bajos con medianas cercanas al 0, lo que supone una falta de inversión y un incremento en riesgos de contenido dañino publicado en estos idiomas.

Estos datos nos muestran un desequilibrio en la distribución de moderadores en los diferentes idiomas de la Unión Europea, de la que queda completamente excluido, de partida, el idioma maltés, idioma

oficial en la organización. La distribución desigual de recursos en las redes plantea preguntas sobre la equidad en la moderación, esto puede suponer que algunas comunidades de idiomas con bajos promedios puedan estar más expuestas a contenidos nocivos por falta de supervisión. Por otro lado la alta dependencia de ciertos idiomas hacia redes específicas constituye un alto riesgo sistémico, pues el cambio de políticas en las moderaciones podría suponer dejar sin moderación adecuada a comunidades enteras.

7. Conclusiones

La Comisión de la Unión Europea se pronunció sobre los informes, analizados en este artículo, presentados en el ejercicio del segundo cuatrimestre del año 2023 con relación al cumplimiento de la ley DSA, reportando tan solo una denuncia hacia la red X por falta de transparencia en los datos y carencia de verificación de las cuentas. Además se inició una investigación a la red META por exceso de publicidad engañosa, y a Tik Tok se le abrió una investigación sobre la protección de los menores, por su diseño adictivo generado en su aplicación Tik Tok Lite, que incentivaría a los menores a utilizar la herramienta a cambio de recompensas. (Wittek, 2024)

Atendiendo al dictamen de la comisión entendemos que el papel de las plataformas con relación a la gestión de recursos de moderación de contenidos es totalmente correcto, a pesar del desequilibrio que hemos manifestado en este estudio, sin considerar las implicaciones posibles en las diferentes comunidades.

La inversión de recursos humanos destinada a la moderación de contenidos no solo debe ser significativa, sino también estratégicamente distribuida para cubrir las necesidades específicas de los usuarios en términos de idioma y cultura.

En la Unión Europea existe 24 lenguas oficiales, y aunque la Ley de Servicios Digitales, no refleja de manera explícita la necesidad de cubrir la moderación de contenidos íntegramente todas las lenguas oficiales, el organismo, si apoya el multilingüismo y considera la necesidad del reconocimiento y defensa de las lenguas, que permiten una mayor competitividad y cohesión social, así lo defiende la Carta de los derechos fundamentales de la UE (2000).

Un enfoque equitativo implicaría que el número de usuarios se ajusta proporcionalmente al volumen de moderadores, al volumen de los contenidos y a la diversidad de idiomas. Es imprescindible asegurar la representación lingüística para proteger los derechos fundamentales de los ciudadanos y garantizar la experiencia inclusiva en las plataformas. La exclusión de una lengua podría derivar en una desprotección sistemática de sus hablantes dejado un vacío crítico en la capacidad de atender denuncias o regular contenidos nocivos. Esta circunstancia no solo pone en riesgo a los usuarios de las plataformas sino que puede suponer una amenaza en la cohesión social y una falta de respeto a la diversidad cultural. Sería imprescindible analizar las implicaciones éticas y sociales consecuentes de la desigualdad en la asignación de moderadores, especialmente en contextos de lenguas minoritarias.

La razón de proporción inadecuada de moderadores por usuarios puede además tener consecuencias graves en la retirada de material perjudicial, que puede verse ralentizada o incluso puede pasar desapercibida ocasionando una falta de control que genere un entorno digital más hostil. Las cifras entregadas en los informes de transparencia en el cumplimiento de la Ley DSA, son insuficientes para evaluar si esta proporción es adecuada y si realmente las empresas tecnológicas están asumiendo la responsabilidad de invertir de manera justa y estratégica en la moderación de contenido.

Estas cifras son también, deficientes para analizar realmente el volumen de trabajo real que cada moderador tiene, y el impacto de la naturaleza de los contenidos, dato que sería imprescindible para evaluar la situación real. El desempeño del trabajo bajo condiciones inadecuadas, puede suponer un riesgo psicosocial no contemplados como enfermedades profesionales, según el cuadro reglamentario de enfermedades profesionales, lo que deja desprotegido al trabajador.

La nueva ley debería exigir estándares estructurales en la asignación de recursos humanos y recursos de IA, que protegieran al trabajador, y permitiera una moderación de recursos adecuada que se reflejase en la seguridad de la red respetando además la multiculturalidad.

La moderación de la IA, requiere de supervisión humana como lo demuestra los resultados arrojados por la investigación realizada por la Agencia de los Derechos Fundamentales de la Unión Europea llevado a cabo en noviembre del 2023. En este estudio se analizaron 350.000 publicaciones, concluyeron que de las 1.500 publicaciones previamente evaluadas por herramientas de moderación de contenidos (IA) con aprobación, el 53% recibieron la consideración de discursos de odio por parte de observadores

humanos. El director de la Agencia en la presentación de este estudio pidió reforzar la moderación de contenidos para frenar el odio en redes:

"El gran volumen de odio que identificamos en las redes sociales muestra claramente que la UE, sus Estados miembros y las plataformas digitales pueden intensificar sus esfuerzos para crear un espacio online más seguro para todos y que respete los derechos humanos, incluida la libertad de expresión. Es inaceptable atacar a persona online solo por su género, color de piel o religión" (Vilas, 2023)

Frente a estas conclusiones y precisamente en nombre de la libertad de expresión las plataformas están renunciando a la moderación de contenido, lo que ha suscitado alerta en diferentes foros como el que ha tenido lugar el pasado mes de enero de 2025: El Foro Económico Mundial de Davos, en el que se concluyó que la renuncia de las redes a moderar su contenido conducirá a una mayor erosión de la democracia.

El debate se ha desplazado, bajo el lema libertad de expresión, hacia derroteros peligrosos que pueden provocar no solo esa erosión en la democracia sino también perjuicios graves a nivel individual en el sujeto, y a nivel general en la sociedad. La regulación de la moderación de contenidos es imprescindible para poder garantizar comunicaciones seguras. Las grandes compañías tecnológicas deben invertir en recursos humanos y garantizar condiciones laborales adecuadas que permitan el desarrollo laboral sin malograr la integridad psicosocial del trabajador. De igual modo que es fundamental que las autoridades establezcan legislaciones para regular los contenidos en las redes sociales, ya que esto contribuye a prevenir la desinformación, proteger los derechos de los usuarios y garantizar un entorno digital más seguro y equitativo para todos.

Referencias

Agarwal, A., Mittal, M., Pathak, A., & Goyal, L. M. (2020). Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning. *SN Computer Science*, *1*(3), 143. https://doi.org/10.1007/s42979-020-00165-4

Alharthi, R., Alhothali, A., & Moria, K. (2021). A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter. *Information Systems*, *99*, 101740. https://doi.org/10.1016/j.is.2021.101740

Andročec, D. (2020). Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, *12*(2), 205-216. https://doi.org/10.2478/ausi-2020-0012

Boyd, D & Marwick, A. (2017, enero 4). *Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies*. Paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society". Microsoft Research. https://doi.org/10.31219/osf.io/2gec4

Ceci L. (2023, abril). Statista. *Horas de video subidas a YouTube*. https://es.statista.com/estadisticas/599096/horas-de-video-subido-a-youtube-cada-minuto-2007/

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement with Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1-52. https://doi.org/10.1145/3686967

Committee on Freedom of Access to Information (2020). *IFLA Statement on Libraries and Artificial Intelligence*. International Federation of Library Associations and Institutions (IFLA). https://repository.ifla.org/handle/20.500.14598/1646

Comisión europea (2023) *Paquete de Servicios Digitales* . https://digital-strategy.ec.europa.eu/en/factpages/safer-fairer-online-environment

Dang B., Riedl J., & Lease M. (2018). But Who Protects the Moderators? The Case of Crowdsourced Image Moderation. *arXiv:1804.10999v4 [cs.HC] 5 Jan 2020*. https://doi.org/10.48550/arXiv.1804.10999

Évole Requena J., & Lara, R (Productores). (2024, octubre 13*) Redes Sociales: la fábrica del terror* (Programa de televisión). En Salvados. Atresplayer https://www.atresplayer.com/lasexta/programas/salvados/temporada-21/redes-sociales-la-fabrica-del-terror-parte-2_670949534911b0e4979c8b87/

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Grimmelmann, J. (2017). The Virtues of Moderation. *Yale Journal of Law & Technology*, 42-108. https://doi.org/10.31228/osf.io/qwxf5

Hernández García, V. (2021). Retos de traducción de las «Terms and Conditions» de las redes sociales: análisis jurídico y terminológico contrastivo inglés-español basado en corpus. *Hikma*, *20*(1), 125-156. https://doi.org/10.21071/hikma.v20i1.12938

Jiménez González, A., & Cancela Rodríguez, E. (2023). ¿Es posible gobernar a las plataformas digitales? Análisis crítico de la Ley Europea de Servicios Digitales. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, *20*(1), 91-99. https://doi.org/10.5209/tekn.82074

Lievens, E., Livingstone, S., McLaughlin, S., O'Neill, B., & Verdoodt, V. (2019). Children's Rights and Digital Technologies. En U. Kilkelly & T. Liefaard (Eds.), *International Human Rights* (pp. 487-513). Springer. https://doi.org/10.1007/978-981-10-4184-6_16

Limón, R. (2025, enero 10). Suprimir la verificación y moderación en redes aumenta el odio y el acoso, advierten los expertos | Tecnología | EL PAÍS. *El País*. https://elpais.com/tecnologia/2025-01-10/suprimir-la-verificacion-y-moderacion-en-redes-aumenta-el-odio-y-el-acoso-advierten-los-expertos.html

Livingstone, S., & Third, A. (2017). Children and young people's rights in the digital age: An emerging agenda. *New Media & Society*, *19(5),* 657-670. https://doi.org/10.1177/1461444816686318

Lousada Arochena, J. F. (2024). Enfermedades del trabajo en la era digital: ¿están las leyes ajustadas a las nuevas realidades? Los trastornos psiquiátricos de un moderador de contenidos de Internet. *Revista de Jurisprudencia Labora*l, Nº 2/2024. 1-6

Manovich, L. (2020). *Cultural analytics*. The MIT Press.

Mitchell, W. J. T. (2017). Iconology, visual culture and media aesthetics. *Poznańskie Studia Polonistyczne. Seria Literacka*, *30*, 341-364. https://doi.org/10.14746/pspsl.2017.30.17

Newton, C. (2020, mayo 12). Facebook will pay $52 million in settlement with moderators who developed PTSD on the job - The Verge. *www.theverge.com*. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

Nissenbaum, N. (2011). Privacy in Context: Technology, Policy, and Social Life. *German Law Journal. Vol 12.* 957-967 . https://doi:10.1017/S2071832200017168

Núñez-Cansado, M., López-López, A., & Somarriba-Arechavala, N. (2021). Covert advertising by kidsfluencers: A methodological proposal applied to the case study of the ten youngest youtubers with most followers in Spain. *Profesional de la Informacion*, *30*(2). https://doi.org/10.3145/epi.2021.mar.19

Roberts, S. (2019). *Behind the Screen*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Torres Vargas, G. Araceli. (2022). *Desafíos en el entorno de la información y la documentación ante las problemáticas sociales actuales*. Universidad Nacional Autónoma de México.

Unión Europea. (2000). *Carta de los derechos fundamentales de la unión europea*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32000X1219%2801%29

Unión Europea. (2022). *Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo, relativo a los servicios digitales (Ley de Servicios Digitales - DSA) .* Diario Oficial de la Unión Europea.https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32022R2065

Vilas C. (2023, noviembre) *Un informe europeo pide reforzar la moderación de contenidos para frenar el odio en redes sociales. Euronews*. https://www.dailymotion.com/video/x8q2cuy

Wang, K., Fu, Z., Zhou, L., & Zhu, Y. (2022). Content Moderation in Social Media: The Characteristics, Degree, and Efficiency of User Engagement. *2022 3rd Asia Symposium on Signal Processing (ASSP)*, 86-91. https://doi.org/10.1109/ASSP57481.2022.00022

Wittek R. (2024, julio).¿Cómo se ha aplicado la Ley de Servicios Digitales durante su primer año en vigor? *Euronews*. https://es.euronews.com/next/2024/07/22/como-se-ha-aplicado-la-ley-de-servicios-digitales-durante-su-primer-ano-en-vigor

X Corp. (2024). *DSA Transparency Report - October 2024*. https://transparency.x.com/dsa-transparency-report.html

Zuckerberg, M. (2025). *It's time to get back to our roots around free expression [Video]*. https://www.facebook.com/watch/?v=1525382954801931