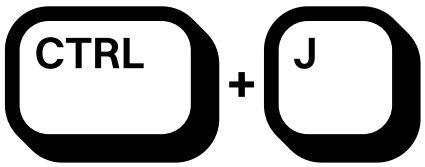


THE PROTOCOL GAP: BRASIL Marzo 2026

Momentum
Journalism & Tech Task Force

International Fund
for Public Interest Media

Journalism
Relay
Project

Este informe es una publicación conjunta de Journalism Relay Project, Momentum – Journalism and Tech Task Force y del International Fund for Public Interest Media (IFPIM).

Sérgio Spagnuolo (Journalism Relay Project) desarrolló la metodología y lideró el análisis de datos. Ester Borges, Bruno Fiaschetti y Paula Miraglia (Momentum) aportaron el análisis contextual para Brasil. Jessica White y Irene Jay Liu (IFPIM) brindaron apoyo general en edición e investigación.

Contenidos

The Protocol Gap: Brasil

Principales hallazgos

- Una abrumadora mayoría de los sitios de noticias en Brasil no tienen política alguna sobre el rastreo por inteligencia artificial (IA) o adopta un enfoque muy permisivo. Solo el 7,2% de los sitios de noticias brasileños prohíben al menos un rastreador de IA mediante un archivo robots.txt, a pesar de que la mayoría de los sitios de noticias (75%) cuenta con un archivo robots.txt en su sitio web.
- Los sitios web en Brasil que implementan directivas en robots.txt concentran sus esfuerzos en los crawlers de IA más conocidos de OpenAI, Common Crawl, Google, ByteDance, Amazon, Apple, Meta y Huawei.
- Estos datos indican que muchos sitios web no utilizan robots.txt como un mecanismo para señalar sus preferencias respecto a que las empresas de IA rastreen su contenido.
- Mantener un archivo robots.txt actualizado no es una solución infalible, pero sí una herramienta fácilmente disponible para que las redacciones indiquen sus preferencias respecto al uso de su contenido, en coherencia con la estrategia y el modelo de negocio de la organización. Sea cual sea su postura, las organizaciones necesitan ofrecer una orientación pública clara sobre cómo se utiliza su contenido para entrenar nuevos modelos de IA.

Introducción

En los últimos años, los sitios de noticias han empezado a recibir un nuevo tipo de visitante: los crawlers (rastreadores) de IA. Desde Googlebot hasta el "GPTbot" de OpenAI, el "Claudebot" de Anthropic y el "ExternalAgent" de Meta, todos estos crawlers tienen algo en común: una necesidad persistente de datos para entrenar grandes modelos de lenguaje (LLM), desarrollar asistentes de IA y alimentar motores de búsqueda impulsados por IA. Estos nuevos visitantes llaman la atención especialmente si consideramos que cerca del 30% del tráfico web global hoy proviene de bots. Los sitios de noticias son destinos atractivos porque producen información oportuna y confiable que puede mejorar la calidad de los modelos y resultados de IA.

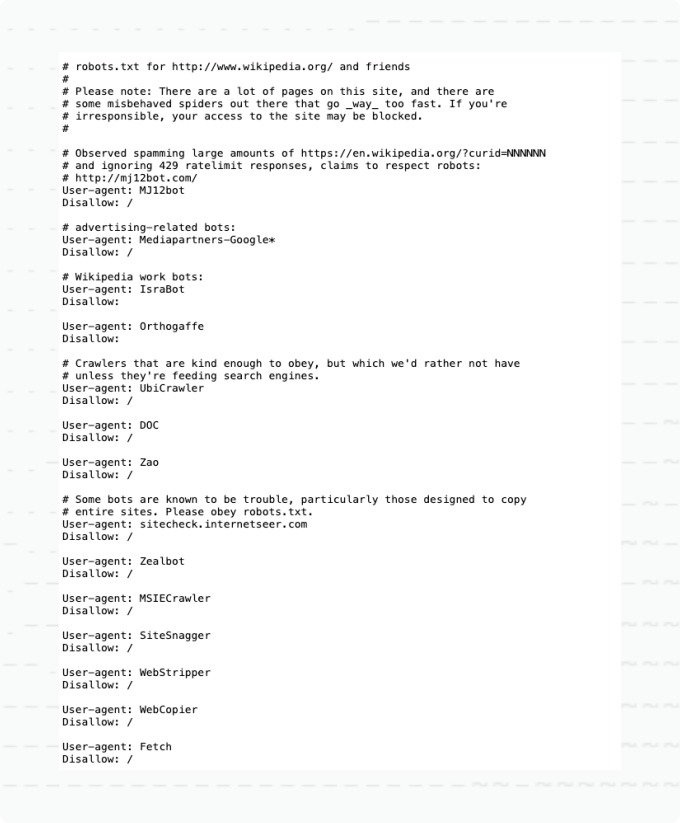
Esta creciente demanda de datos para IA presenta nuevas oportunidades—pero también numerosos desafíos—para las organizaciones mediáticas. Genera importantes cuestiones legales y éticas en torno al uso no autorizado, la infracción de derechos de autor y las violaciones de privacidad. Pero también pone en juego la propia supervivencia del periodismo: sin mecanismos para señalar y hacer cumplir permisos a los crawlers de IA, las organizaciones mediáticas no pueden proteger ni monetizar eficazmente el valor de su trabajo cuando es rastreado, reutilizado y redistribuido por sistemas de IA. Esta brecha corre el riesgo de acelerar la crisis de sostenibilidad que enfrentan los medios, ya afectados por la caída de visibilidad y tráfico procedente de las plataformas digitales—lo que muchos editores ya llaman un "apocalipsis del tráfico"—y por el colapso de los modelos tradicionales basados en publicidad.

Robots.txt es un archivo de texto añadido al dominio raíz de un sitio web que contiene instrucciones para los motores de búsqueda y los crawlers web sobre qué páginas o archivos pueden acceder. A pesar de sus limitaciones, robots.txt es una de las pocas herramientas gratuitas disponibles para que los creadores de contenido indiquen si desean que sus datos sean utilizados por empresas de IA. Robots.txt ya ha sido utilizado como evidencia en demandas contra el rastreo y scraping no autorizados en Canadá, Estados Unidos y el Reino Unido. El archivo también puede servir como elemento de negociación en acuerdos de licenciamiento, especialmente en contextos en los que los modelos de IA dependen en gran medida de grandes volúmenes de datos no consentidos y no remunerados.

Los sitios de noticias son cada vez más conscientes de este desafío creciente, pero actualmente existen pocas formas de gestionar a los crawlers de IA y ninguna de ellas es completamente infalible. Algunos sitios están implementando muros de pago estrictos, mientras que otros emprenden extensas demandas por infracciones de derechos de autor. También están surgiendo respuestas técnicas: en julio de 2025, el proveedor de infraestructura y seguridad web Cloudflare anunció que bloquearía por defecto los crawlers de IA y ayudaría a aplicar un modelo basado en permisos. Mercados de contenido—como el piloto lanzado por Microsoft

o startups como ProRata.ai y Tollbit—y nuevas soluciones de datos desarrolladas por empresas como Flexolmo y OpenMined también emergen como formas de dar a los productores de contenido mayor control sobre cómo se accede, cómo se utiliza y cómo se compensa su contenido por parte de las empresas de IA. Pero muchos medios aún recurren al Protocolo de Exclusión de Robots (robots.txt) para gestionar el raspado (scraping) de contenido por parte de crawlers de IA.

A pesar de la creciente conciencia sobre el rastreo y scraping no autorizados para IA, las respuestas de los medios para proteger su contenido en línea siguen siendo, en el mejor de los casos, fragmentarias. Los medios pequeños y medianos se encuentran en una desventaja mayor porque no necesariamente cuentan con los recursos para invertir en infraestructura sofisticada que detecte y bloquee crawlers de IA, ni tienen el peso para negociar acuerdos de licenciamiento individuales con grandes empresas de IA: Reddit, por ejemplo, acordó licenciar su contenido a Google para entrenamiento de IA en febrero de 2024.



```
# robots.txt for http://www.wikipedia.org/ and friends
#
# Please note: There are a lot of pages on this site, and there are
# some misbehaved spiders out there that go _way_ too fast. If you're
# irresponsible, your access to the site may be blocked.
#

# Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN
# and ignoring 429 ratelimit responses, claims to respect robots:
# http://mj12bot.com/
User-agent: MJ12bot
Disallow: /

# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /

# Wikipedia work bots:
User-agent: IsraBot
Disallow:

User-agent: Orthogaffe
Disallow:

# Crawlers that are kind enough to obey, but which we'd rather not have
# unless they're feeding search engines.
User-agent: UbiCrawler
Disallow: /

User-agent: DOC
Disallow: /

User-agent: Zao
Disallow: /

# Some bots are known to be trouble, particularly those designed to copy
# entire sites. Please obey robots.txt.
User-agent: sitecheck.internetseer.com
Disallow: /

User-agent: Zealbot
Disallow: /

User-agent: MSIECrawler
Disallow: /

User-agent: SiteSnagger
Disallow: /

User-agent: WebStripper
Disallow: /

User-agent: WebCopier
Disallow: /

User-agent: Fetch
Disallow: /
```

Captura de pantalla del archivo robots.txt utilizado por Wikipedia.org

Sobre este proyecto

Este informe es el resultado de una colaboración entre Journalism Relay Project, Momentum – Journalism and Tech Task Force, y el International Fund for Public Interest Media. En este primer informe por país centrado en Brasil, buscamos mapear cuántos sitios de noticias utilizan robots.txt para gestionar crawlers de IA. Forma parte de un proyecto de investigación más amplio que involucra socios en Brasil, Indonesia y Sudáfrica, y que tiene como objetivo fortalecer nuestra comprensión global de las normas y prácticas en evolución sobre el acceso de la IA al contenido periodístico, con un enfoque particular en los mercados clave del Sur Global.

A través de investigaciones técnicas y cualitativas, buscamos aumentar la conciencia entre las organizaciones mediáticas sobre cómo los sistemas de IA acceden a su contenido y ayudar a informar estrategias y políticas para gestionar los crawlers de IA de manera que mejor sirvan a los intereses de las redacciones. Hemos desarrollado una metodología detallada (ver anexo de este informe) para todos aquellos que deseen replicar esta investigación sobre robots.txt en otros países y regiones.

Si bien robots.txt sigue siendo el mecanismo más ampliamente disponible, esta investigación comienza a revelar brechas críticas en el uso de robots.txt por parte de las redacciones para gestionar crawlers de IA, con un enfoque en los mercados del Sur Global. En este primer informe, que analiza más de 4.000 sitios de noticias en Brasil, encontramos que la gran mayoría no da ninguna directiva respecto a los crawlers de IA, aunque muchas organizaciones mediáticas ya cuentan con un archivo robots.txt adjunto a su sitio.[1]

A medida que las empresas de IA continúan luchando por la dominancia del mercado y recolectando enormes volúmenes de datos para alimentar sus sistemas, la necesidad de respuestas más informadas y coordinadas es especialmente crucial para las organizaciones mediáticas del Sur Global. La adopción coordinada de robots.txt entre los medios puede ayudar a establecer una participación compartida en estándares que beneficien a toda la comunidad periodística—desde grandes medios hasta editores independientes—al establecer directivas claras hacia las empresas de IA. Estos sencillos documentos de texto siguen siendo una herramienta relevante para empoderar a las organizaciones mediáticas en la era de la IA, transformando lo que de otro modo sería una extracción pasiva de datos en una decisión editorial activa.

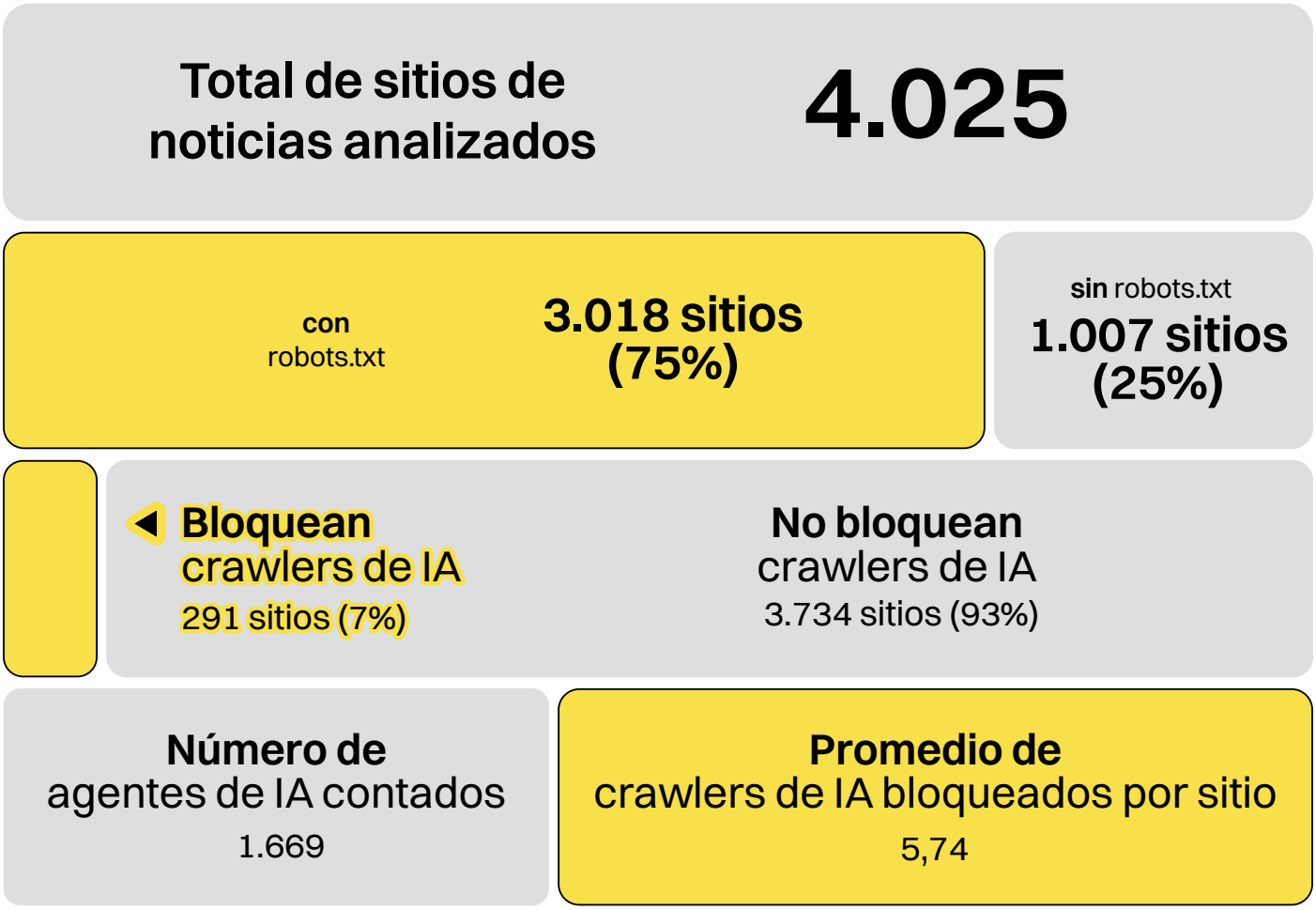
[1]Los datos de los sitios web de Brasil se basaron en la investigación nacional más reciente de Atlas da Notícia (atlas.jor.br), de junio de 2025.

Robots.txt: una herramienta poco utilizada en Brasil

Nuestro análisis encontró que una abrumadora mayoría de los sitios de noticias en Brasil no tienen ninguna política sobre el rastreo por IA o adopta un enfoque muy permisivo, y varios sitios muestran únicamente los archivos robots.txt básicos de sus sistemas de gestión de contenido, si es que tienen alguno.

Apenas alrededor del **7,2%** de los sitios de noticias brasileños tenían al menos una directiva para IA, aunque la mayoría (**75%**) contaba con un archivo robots.txt adjunto a su sitio web.

Robots.txt entre los editores brasileños: lo que nos dicen los datos



Fuente: The Protocol Gap

Esta es una fuerte indicación de que la gran mayoría de los sitios web desconocen las capacidades de establecer declaraciones públicas sobre el grado de permisividad frente al rastreo de su contenido por parte de IA, y en cambio confían en el archivo robots.txt predeterminado que viene por defecto en sus sistemas de gestión de contenido.

Los sitios web en Brasil que bloquean IA adoptan un enfoque mucho más específico, concentrando sus esfuerzos en lo que parecen ser los crawlers de IA más prominentes en lugar de aplicar un bloqueo generalizado. Los sitios brasileños bloquean en promedio solo cuatro agentes de IA por sitio que realiza bloqueo.

Esta estrategia de bloqueo selectivo podría reflejar varios factores: estos países podrían tener menos conciencia sobre el espectro completo de crawlers de IA que operan en línea, sus sitios web podrían estar principalmente preocupados por bloquear solo los sistemas de IA más conocidos, o podrían existir diferencias culturales y regulatorias que hacen que el bloqueo completo de IA sea una prioridad menor en comparación con el enfoque más cauteloso adoptado por las organizaciones mediáticas en países como Estados Unidos. Se requiere un estudio adicional para comprender los factores predominantes.

Agentes bloqueados

Entre los sitios web que bloquean al menos un agente de IA, los 10 bots más bloqueados son:

Agente de IA	Descripción	% de bloqueos*
gptbot	De OpenAI. Los datos se utilizan para entrenar modelos actuales y futuros, eliminando datos con paywall, PII y datos que violan las políticas de la empresa.	10,2
ccbot	De Common Crawl Foundation. Archivo web que se remonta a 2008. Citado en miles de artículos de investigación por año.	9,77
google-extended	De Google. Utilizado para entrenar las APIs generativas de Gemini y Vertex AI. No afecta la inclusión ni la clasificación de un sitio en la Búsqueda de Google.	9,53
claudebot	De Anthropic. Extrae datos para entrenar LLMs y productos de IA ofrecidos por Anthropic.	9,17
bytespider	De ByteDance. Descarga datos para entrenar LLMs, incluidos competidores de ChatGPT.	8,81
amazonbot	De Amazon. Incluye referencias a sitios rastreados cuando se muestran respuestas vía Alexa; no detalla claramente otros usos.	8,51
applebot-extended	De Apple. Apple tiene un agente secundario, Applebot-Extended ... [que] se utiliza para entrenar los modelos fundacionales de Apple que impulsan las funciones generativas de IA en productos Apple, incluyendo Apple Intelligence, Servicios y Herramientas para Desarrolladores.	8,39
meta-externalagent	De Meta. Meta-ExternalAgent rastrea la web para casos de uso como entrenar modelos de IA o mejorar productos mediante la indexación directa de contenido.	8,21
petalbot	De Huawei. Operado por Huawei para proporcionar servicios de búsqueda y asistentes de IA.	3,82
universal_block	Utiliza la directiva para bloquear todos los agentes externos. User-agent: * Disallow: /	1,92

Para ver una lista de autorías y descripciones de bots, consulte: https://github.com/ai-robots-txt/ai.robots.txt/blob/main/table-of-bot-metrics.md

* Porcentaje calculado a partir de todos los bloqueos observados (no es el porcentaje de sitios web de noticias que bloquean).

Implicaciones para los editores en el cambiante panorama de la IA en Brasil

El uso limitado de robots.txt para desautorizar crawlers de IA pone de manifiesto una brecha entre los editores de noticias brasileños mientras enfrentan el impacto de las tecnologías emergentes. Las redacciones en Brasil están adoptando cada vez más herramientas de IA para una variedad de propósitos, desde traducir artículos hasta producir videos. Pero el auge de la IA también está profundizando los desafíos de sostenibilidad. En un mercado mediático históricamente concentrado en unos pocos grandes conglomerados, la falta de respuestas proactivas y coordinadas corre el riesgo de agravar desafíos relacionados con relaciones desiguales con las plataformas digitales y dificultades para negociar colectivamente con las grandes empresas tecnológicas.

Muchos editores brasileños han expresado preocupación por el uso no autorizado de su contenido, los riesgos éticos y la disminución del tráfico web a la vez que manifiestan un deseo de una regulación y acuerdos justos, pero aún no han surgido estrategias colectivas, ni acuerdos concretos. Hasta la fecha, no se han hecho públicos acuerdos de licencia entre organizaciones de medios y plataformas tecnológicas en Brasil. En agosto de 2025, Folha de São Paulo, uno de los mayores periódicos del país, presentó una demanda acusando a OpenAI de competencia desleal e infracción de derechos de autor, alegando que la empresa utilizó su contenido sin permiso ni pago para entrenar sus modelos de IA. AI Overviews (Google) también está intensificando las presiones financieras, con evidencia que muestra su impacto en el tráfico hacia los sitios de noticias y en posibles fuentes de ingresos.

No proteger el contenido puede representar una oportunidad perdida para asegurar tanto recursos como legitimidad para el periodismo. Desde una perspectiva de sostenibilidad, la remuneración por el uso de contenido periodístico podría ser una fuente vital de financiación para compensar, al menos parcialmente, la pérdida de audiencia e ingresos publicitarios resultante del cambio digital en el consumo de información. Frente a tensiones financieras continuas, el ecosistema mediático podría encontrar en esta nueva fuente de financiación una forma de fortalecer su papel democrático.

Desde el punto de vista de la legitimidad, la extracción de contenido periodístico sin reconocimiento de su valor social puede erosionar aún más la posición del periodismo como fuente primaria de información. En Brasil, estudios recientes revelan que, particularmente entre las poblaciones más jóvenes, las plataformas de redes sociales se han convertido en las principales herramientas para acceder a información, incluidas noticias y contenido político.

Agravando este problema está el hecho de que la mayoría de estas empresas y sus tecnologías tienen sede en Estados Unidos y China—lo que significa que, tanto en su diseño, como en su operación, estas herramientas no solo son inadecuadas para satisfacer las necesidades y especificidades del Sur Global, sino que a menudo refuerzan y profundizan las asimetrías de poder Norte-Sur.

A pesar de los muchos desafíos, Brasil se encuentra en un momento crucial para actuar. Los tres poderes del Estado—Ejecutivo, Legislativo y Judicial—están avanzando en medidas que podrían redefinir el futuro del periodismo.

El gobierno federal de Brasil, en consonancia con las aspiraciones más amplias de otros países del Sur Global, ha pedido mayor protagonismo en los debates globales sobre soberanía digital y regulación de la IA. Ejemplos recientes incluyen la declaración del Grupo de Trabajo de Economía Digital del G20 sobre integridad de la información en 2024, así como la más reciente declaración de los BRICS presentada en la cumbre de Río de Janeiro en julio de 2025, que situó a la IA como una alta prioridad en la agenda del bloque. El compromiso del Ejecutivo podría acelerar el progreso regulatorio y abrir espacio para que la sociedad civil y el sector periodístico den forma a estas discusiones.

Mientras tanto, el poder legislativo está avanzando en proyectos clave orientados a regular las plataformas y su relación con el periodismo:

- Proyecto de Ley N.º 2630 (2020) busca equilibrar la libertad de expresión y la lucha contra la desinformación, exigiendo mayor transparencia algorítmica y obligando a las plataformas que se beneficien del contenido periodístico a compensar a quienes lo producen.

- Proyecto de Ley de IA – N.º 2338 (2023) establece directrices generales para el desarrollo y uso de la IA en Brasil. Su objetivo es garantizar prácticas responsables que protejan derechos y sistemas democráticos. En cuanto a la actividad periodística específicamente, el proyecto establece que el uso automatizado de obras—como extracción, reproducción, almacenamiento y transformación, incluida la minería de datos y textos en sistemas de IA—por instituciones de investigación, organizaciones de medios, museos, archivos y bibliotecas, no constituye infracción de derechos de autor. Esta exención depende de que las herramientas se utilicen solo lo necesario, sin perjudicar los intereses económicos de los titulares de los derechos, sin competir con la explotación normal de las obras y sin fines comerciales sin ofrecer una compensación justa. La discusión sobre la obligación de remunerar por el uso de contenido protegido es, sin embargo, el punto más sensible del proyecto. Las grandes empresas tecnológicas están presionando a los diputados para evitar el pago, lo que hace que los legisladores sean cautelosos a la hora de avanzar en regulaciones más estrictas, dadas las implicaciones económicas y geopolíticas de tales medidas.

El Supremo Tribunal Federal también ha moldeado este debate a través de iniciativas como la "Investigación de las Fake News" y la decisión de junio de 2025 que redefinió la responsabilidad de las redes sociales por contenido generado por usuarios. Aunque la postura del tribunal ha generado preocupaciones sobre el riesgo de sobrerregulación y su impacto en la libertad de expresión, ha influido en el debate nacional sobre gobernanza y regulación de la IA.

En medio de estos cambios políticos, la industria de noticias sigue siendo en gran medida reactiva ante las convulsiones tecnológicas. Este estudio podría servir como un llamado de atención. Con solo el 7,2% de los medios brasileños utilizando actualmente robots.txt para bloquear crawlers de IA, los hallazgos resaltan la urgencia de profundizar las discusiones sobre la apropiación de contenido y desarrollar vías viables para su reconocimiento y remuneración. Fortalecer la coordinación del sector en torno a estos temas podría representar el primer paso hacia un futuro digital más sostenible para el periodismo brasileño.

Donde robots.txt queda corto, las redacciones pagan

Porque Google no distingue entre su motor de búsqueda tradicional y su función de búsqueda AI Overviews, los editores no pueden usar fácilmente el archivo robots.txt para evitar que su contenido sea utilizado en resúmenes generados por IA sin desaparecer por completo de los resultados de búsqueda. Esta disyuntiva conlleva consecuencias económicas significativas. Muchos editores —incluidos los de Brasil— han reportado fuertes caídas en el tráfico a sus sitios web desde que se lanzó AI Overviews, lo que está afectando su capacidad de generar ingresos. Esto está relacionado con que los usuarios son menos propensos a hacer clic en los sitios de noticias si obtienen la información a través de un resumen generado por IA. En septiembre de 2025, Cloudflare anunció una Política de Señales de Contenido (Content Signals) que se integraría en un archivo robots.txt y podría dar a los propietarios de sitios un control más sólido para bloquear crawlers de entrenamiento de IA mientras siguen siendo visibles en la búsqueda tradicional. Sin embargo, aún está por verse si esto detendrá la pérdida de tráfico web.

Conclusión y próximos pasos

Este estudio sirve como punto de partida para reflexiones más profundas sobre la compleja relación entre el periodismo y las plataformas digitales en el Sur Global—particularmente en lo que respecta al uso no autorizado de contenido periodístico por parte de herramientas de IA generativa y al desarrollo de estrategias que permitan compensar de manera justa a los medios de comunicación.

A pesar de sus limitaciones—sobre todo, las restricciones técnicas de robots.txt como herramienta de protección de contenidos y la dependencia exclusiva de datos públicos—los hallazgos señalan una sensación más amplia de falta de preparación, o incluso de desesperanza, entre los editores brasileños ante la captura de su contenido por parte de empresas de IA. La toma de conciencia sobre cuán pocos medios brasileños protegen activamente su contenido mediante robots.txt también podría marcar un punto de inflexión, alejándose de la fragmentación y la falta de participación hacia una discusión colectiva sobre la regulación de la IA, además de inspirar una nueva ola de reflexión estratégica dentro de la industria de medios.

Este estudio representa el primer capítulo de una colaboración entre Journalism Relay Project, Momentum – Journalism and Tech Task Force, y el International Fund for Public Interest Media (IFPIM), con el objetivo de construir un cuerpo crítico de conocimiento y contribuir a las discusiones sobre caminos para la sostenibilidad del periodismo desde la perspectiva del Sur Global.

Los próximos pasos incluyen:

1. **Investigación adicional:** Con base en los datos y conclusiones presentados aquí, planeamos avanzar hacia un análisis cualitativo del panorama brasileño mediante entrevistas con editores locales. Nuestro objetivo es recopilar nuevas percepciones que nos ayuden a comprender de forma más clara y profunda los dilemas que enfrenta el periodismo en Brasil en relación con la captura de contenido y la búsqueda de marcos regulatorios justos para el uso de su material.

2. **Herramientas para editores:** La decisión de bloquear o no a los crawlers es una decisión individual que debe ajustarse a la estrategia de cada medio.

3. **Cooperación transfronteriza:** La colaboración también busca replicar la metodología utilizada en este estudio para examinar los ecosistemas mediáticos de Indonesia y Sudáfrica, otros dos países del Sur Global que, al igual que Brasil, comparten características similares: mercados mediáticos históricamente concentrados, relaciones desiguales con las plataformas digitales y dificultades para negociar colectivamente con las grandes empresas tecnológicas.

Invitamos a todas las personas investigadoras que deseen replicar este estudio en otros mercados a consultar la metodología completa que se presenta a continuación.

Metodología

1.1.Origen de los datos

Este proyecto utilizó verificación programática de archivos para determinar la existencia de parámetros de bloqueo para agentes de IA en archivos robots.txt de 4.023 organizaciones de noticias de Brasil. Los datos de sitios web en Brasil se basaron en la investigación nacional más reciente del Atlas da Notícia (atlas.jor.br), de junio de 2025. Las principales variables consideradas fueron el nombre de las organizaciones periodísticas, sus direcciones web y sus ubicaciones.

La metodología se centró en examinar estos archivos en busca de patrones específicos y su relación con crawlers de IA, basándose en el parámetro estándar **Disallow:/** en archivos robots.txt. Mediante la consulta de una base de datos de código abierto con identificadores de crawlers de IA conocidos—incluidos los de grandes empresas tecnológicas como Anthropic, Amazon, Apple, Google, Meta y OpenAI—el análisis identificó cuándo los sitios web intentaban específicamente señalar restricciones a los crawlers de IA respecto de su contenido.

Este aspecto de la metodología fue particularmente relevante debido a las preocupaciones recurrentes sobre sistemas de IA que extraen contenido web sin permiso explícito. Es importante mencionar que los archivos robots.txt son instrumentos no vinculantes que no pueden impedir técnicamente la recolección de datos por cualquier bot, sino que únicamente indican la política del sitio frente a la agregación automatizada de contenido.

Al examinar un sitio web, el sistema primero aseguró el formato adecuado de la URL antes de intentar acceder a su archivo robots.txt. Luego implementó protocolos de manejo de errores y límites de tiempo para gestionar distintos escenarios reales, como tiempos de espera del servidor, archivos faltantes, problemas de conexión o bloqueos de DNS, creando un sistema integral de registro de eventos.

Para considerar servidores de respuesta lenta que pudieran necesitar tiempo para “despertar”, el sistema implementó un mecanismo de reintento con un retraso de 3 segundos cuando los archivos robots.txt no se encontraban en el primer intento. Este enfoque mejoró significativamente la fiabilidad de la detección, manteniendo al mismo tiempo tiempos de procesamiento eficientes en servidores con buena respuesta.

La lógica de análisis distinguió entre directivas restrictivas (**Disallow: /** o **Disallow: ***) y directivas permisivas (**Allow:** o **Disallow:** vacío), garantizando una clasificación precisa de las políticas de bloqueo de IA.

Los sitios fueron categorizados en un espectro que iba de permisivos a súper restrictivos, según la cantidad y el tipo de agentes de IA que bloqueaban.

Los sitios con archivos robots.txt malformados o incompletos fueron generalmente clasificados como permisivos hacia los crawlers de IA. Esto incluyó sitios con directivas **Disallow:** vacías (que técnicamente permiten el acceso) y aquellos que utilizaban solo reglas parciales de bloqueo, restringiendo URLs o directorios específicos mientras dejaban la mayor parte del contenido sin restricciones. En ausencia de directivas de bloqueo completas, estos sitios fueron considerados como permitiendo efectivamente el acceso de IA a su contenido.

Se verificaron manualmente 90 resultados aleatorios y se confirmó que la mayoría eran precisos. A pesar de estas medidas metodológicas robustas, por precaución estimamos que persiste una tasa de error del 3% debido a factores técnicos, incluidos problemas de resolución DNS, tiempos de espera del servidor y fallas temporales de conectividad que pueden generar falsos positivos o, en algunos casos, falsos negativos en la detección de robots.txt.

1.2 Recolección de datos

Una vez que el contenido de robots.txt fue recuperado con éxito, el análisis avanzó al reconocimiento de patrones. El sistema procesó el contenido del archivo línea por línea, identificando las directivas User-agent y examinándolas en comparación con la base de datos conocida de crawlers de IA (buscando el parámetro **/Disallow**).

Esta comparación se implementó sin sensibilidad a mayúsculas y minúsculas, para detectar variaciones en la forma en que podrían escribirse los nombres de los crawlers, reduciendo redundancias. Los resultados se estructuraron para ofrecer información sobre la postura de cada sitio frente a los crawlers de IA.

Para cada sitio analizado, el sistema determinó si existía un archivo robots.txt, si contenía directivas de bloqueo y, específicamente, qué crawlers de IA conocidos estaban bloqueados mediante el parámetro **/Disallow**.

Este enfoque estructurado de recolección y análisis permitió realizar de manera rápida (un promedio de 90 minutos de ejecución de código) un examen a gran escala sobre cómo diferentes sitios web abordan el acceso de crawlers de IA.

La recolección y el análisis se realizaron en el lenguaje de programación R.

[2]Atlas da Notícia, sitio - https://atlas.jor.br
[3]Lista de crawlers de IA conocidos, 2025, repositorio de Github – ai.robots.txt

[4]OECD (2025), “Intellectual property issues in artificial intelligence trained on scraped data”, OECD Artificial Intelligence Papers, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/d5241a23-en.

1.3 Clasificación de datos

A medida que los datos eran recolectados de los archivos robots.txt, el algoritmo generaba automáticamente clasificaciones basándose en el nivel de permisividad asignado discretamente a cada sitio.

En lugar de utilizar umbrales absolutos arbitrarios, el sistema primero calcula el promedio base del número de agentes de IA bloqueados por los sitios que efectivamente implementan políticas de bloqueo. Este promedio sirve como punto de referencia para determinar qué constituye un comportamiento de bloqueo normal, moderado o excesivo dentro del conjunto específico de datos analizado.

El sistema de clasificación crea luego cinco niveles distintos utilizando múltiplos de este promedio calculado.

Este enfoque metodológico garantiza que la clasificación siga siendo significativa independientemente de las cifras absolutas de cada conjunto de datos, ya que se adapta a la distribución real de los comportamientos de bloqueo observados. Por ejemplo, si el promedio de bloqueo de un sitio es de tres agentes de IA, un sitio que bloquee treinta agentes sería clasificado como súper restrictivo; pero si el promedio del conjunto de datos fuera de diez agentes, ese mismo sitio que bloquea treinta agentes pasaría a la categoría restrictiva.

Clasificación	Agentes de IA bloqueados	Descripción
Política permisiva hacia agentes de IA	0 agentes de IA	Sitios que no imponen ninguna restricción a los crawlers de IA
Poco restrictiva	1x a 2x el promedio	Sitios con un comportamiento de bloqueo típico o ligeramente superior al promedio
Restrictiva	2x a 5x el promedio	Sitios que bloquean entre dos y cinco veces el promedio
Muy restrictiva	5x a 10x el promedio	Sitios que bloquean entre cinco y diez veces el promedio
Súper restrictiva	Más de 10x el promedio	Sitios con políticas de bloqueo excepcionalmente completas

1.4 Agregación de datos

Finalmente, los datos se agregaron en un conjunto de indicadores que pudieran ser representativos del conjunto completo del dataset.

Metrica	Valor / unidad	Datos observados
total_observations	data / websites	data
block_ai	data / websites	data
not_block_ai	data / websites	data
pct_not_blocked	data / percent	data
pct_blocked	data / percent	data
yes_robots_txt	data / websites	data
no_robots_txt	data / websites	data
pct_robots_txt	data / percent	data
pct_no_robots_txt	data / percent	data
ai_agents_count	data / ai agents	data
avg_agents_per_blocking_site	data / ai agents	data

/members/api/comments/counts/ Disallow: /r/ Disallow: /webmentions/receive/ # Allow rules User-agent: Google Disallow rules # robots.txt generated by Robots.txt Helper AI2Bot Disallow: / User-agent: Ai2Bot-Dolma Disallow: / U aiHitBot Disallow: / User-agent: Amazonbot Disallow: / Use Andibot Disallow: / User-agent: anthropic-ai Disallow: / Us Applebot Disallow: / Use -agent: Applebot-Extended Disal User-agent: bedrockbot Disallow: / User-agent: Brightbot User-agent: Bytespider Disallow: / User-agent: CCBot Disa User-agent: ChatGPT-User Disallow: / User-agent: Claude Disallow: / User-agent: Claude-User Disallow: / User-age Disallow: / User-agent: THE PROTOCOL GAP: ClaudeBot D User-agent: cohere-ai Disallow: / User-agent GoogleOthe Disallow: / User-agent:cohere-training-data-crawler Disa User-agent: Cotoyogi Disallow: / User-agent: Crawlspace User-agent: BRAZIL Disallow: / User-agent: DuckAssistBo User-agent: EchoboxBot Disallow: / User-agent: Faceboo / User-agent: Factset_spyderbot Disallow: / User-agent: F Disallow: / User-agent: FriendlyCrawler Disallow: / User-a Google-CloudVertexBot Disallow: / User-agent: Google-Ex Disallow: / User-agent: GoogleOther Disallow: / User-agen GoogleOther- Marzo 2026 : / User-agent: GoogleOther-Vi User-agent: GPTBot Disallow: / User-agent: iaskspider/2.0 User-agent: ICC-Crawler Disallow: / User-agent: Imagesif / User-agent: img2dataset Disallow: / User-agent: ISSCyb Disallow: / User-agent: Kangaroo Bot Disallow: / User-age meta-externalagent Disallow: / User-agent: Meta-Externa