IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

https://idp.uoc.edu

Monográfico: «Digitalización y algoritmización de la justicia» (coord.: F. Miró)

Particularidades probatorias del discurso de odio en Internet: identificación de indicadores de polarización radical mediante sistemas algorítmicos

Federico Bueno de Mata
Universidad de Salamanca

Fecha de recepción: mayo 2022
Fecha de aceptación: septiembre 2023
Fecha de publicación: noviembre 2023

Resumen

En el presente artículo se aborda la naturaleza jurídica de los indicadores de polarización radical de los delitos de odio como prueba en un proceso penal, y cómo los sistemas algorítmicos pueden detectar su existencia a través de la monitorización y el análisis del discurso de odio en línea, conjugando su uso con técnicas OSINT y con una futurible prueba de inteligencia policial. Así, se parte del encuadramiento de los indicadores de polarización como prueba indiciaria en el proceso para, posteriormente, identificar qué parámetros concretos se deberían tener en cuenta en estos casos para probar el discurso de odio en línea. Finalmente se analizan los algoritmos impulsados a nivel policial en España, planteando sus virtudes y limitaciones legales, y planteando la necesidad de analizar cómo encajan estos algoritmos con la legislación, especialmente en relación con la prohibición de sistemas de policía predictiva y reincidencia criminal.

Palabras clave

odio; OSINT; polarización; indicios; inteligencia

Probative peculiarities of hate speech on the Internet: identification of radical polarization indicators using algorithmic systems

Abstract

This article addresses the legal nature of radical polarization indicators of hate crimes as evidence in a criminal process, and how algorithmic systems can detect their existence by monitoring and analyzing online hate speech, combining their use with OSINT techniques and with future proof of police intelligence. Thus, polarization indicators framing is established as indicative evidence in the process to, subsequently, identify which specific parameters should be taken into account in these cases to prove online hate speech. Finally, the algorithms promoted at a police level in Spain are analyzed, raising their legal virtues and limitations, as well as the need to analyze how these algorithms fit with the legislation, especially in relation to the prohibition of predictive police systems and criminal recurrence.

Keywords

hate; OSINT; polarization; evidence; intelligence

1. La investigación de delitos de odio en redes sociales como objetivo prioritario del Estado

En la actualidad, nos encontramos en una sociedad marcada por la intolerancia y la discriminación hacia ciertos grupos de personas especialmente vulnerables, lo que ha llevado a un recrudecimiento de diversas posturas extremas en distintos sectores. Esta actitud ha dado lugar a un aumento de los delitos de odio, que se han convertido en un tema relevante a nivel judicial, social y político.

En los últimos años, se han realizado esfuerzos para conceptualizar y encajar jurídicamente los delitos de odio en el sistema judicial. Sin embargo, el papel que juegan las nuevas tecnologías en la comisión de este tipo de comportamientos hace que la respuesta penal no sea suficiente para frenarlos. En este sentido, a nivel conceptual España asume la definición ofrecida por el Consejo Ministerial de la Organización para la Seguridad y la Cooperación Europea (OSCE) del delito de odio. Este se entiende como: «toda infracción penal incluidas las cometidas contra las personas o la propiedad, donde el bien jurídico protegido, se elige por su, real o percibida, conexión, simpatía, filiación, apoyo o pertenencia a un grupo. Este grupo se basa en una característica común de sus miembros, como su "raza", real o percibida, el origen nacional o étnico, el lenguaje, el color, la religión, la edad, la discapacidad, la orientación sexual, u otro factor similar».[1] Esta cuestión propiamente penal, que cuenta con un debate doctrinal entre especialistas propios de esta rama de conocimiento,[2] queda fuera de la presente investigación, puesto que partimos de que el delito existe, que tiene su tipificación en el artículo 510 del Código Penal (CP), y que, por ello es necesario combatir esta problemática únicamente desde nuestra disciplina: el derecho procesal.

Concretamente, desde la óptica procesalista es necesario plantearnos si las instituciones procesales existentes deben adaptarse y adecuarse para poder ofrecer una respuesta completa que incluya el tratamiento adecuado de las personas afectadas durante todo el proceso de vic-

1. Decisión n.º 4/03 de la OSCE.
2. Autores como Miró Llinares parten de que no se podrían considerar propiamente como delitos, pues demanda «un replanteamiento de la criminalización de expresiones que parta de la derogación de la mayoría de ofensas que están incluidas en los arts. 578 y 510 CP», concluyendo que gran parte de las mismas deberían solucionarse por vía de ofensas contra el honor, enmarcadas en el Derecho Civil. *Vid*. MIRÓ LLINARES, F. (2017) «Derecho penal y 140 caracteres. Hacia una exégesis restrictiva de los delitos de expresión», *Cometer delitos en 140 caracteres. El Derecho penal ante el odio y la radicalización en Internet*. Madrid: Marcial Pons, pág. 60.

timización, la investigación tecnológica de la comisión de estos hechos en redes abiertas o cerradas, y el posterior tratamiento de las pruebas obtenidas, pues solo así podremos lograr una sociedad más justa e inclusiva, libre de la lacra social que suponen los delitos de odio.

Es así como el derecho procesal afronta esta realidad, como un derecho de garantías y cierre del ordenamiento que se constituye como necesario para proteger los derechos y las libertades fundamentales de los ciudadanos en el espacio cibernético, con el objetivo de consolidar un proceso penal eficaz que pueda hacer frente a una criminalidad cibernética que plantea problemas de base como, la facilidad de viralización unida, la posibilidad de cometer delitos bajo el anonimato o cuentas falsas, y la falta de adaptación normativa a las nuevas realidades tecnológicas, lo que dificulta la investigación y prueba de estas conductas, sin que estas puedan llegar a ser esclarecidas mediante la identificación y la sanción de sus responsables.

Podemos decir que los delitos de odio en Internet representan un desafío considerable para las autoridades policiales y judiciales, ya que su actuación está ligada de manera directa a que desde la ciencia jurídica se articulen mecanismos efectivos para luchar contra esta lacra. España constituye un ejemplo de cómo los Estados deben reaccionar ante una realidad creciente y cada vez más incontrolable, articulando medidas y creando instituciones desde hace más de diez años para intentar que esta ola de odio tecnológico pueda ser refrenada.

En este sentido, en noviembre de 2011, el Consejo de Ministros aprobó la «Estrategia Integral contra el racismo, la discriminación racial, la xenofobia y otras formas conexas de intolerancia», que incluía entre sus objetivos la promoción de mecanismos de detección y protocolos de intervención en caso de incidentes o actitudes racistas, xenófobas o discriminatorias. Con el fin de lograr estos objetivos, se crea la Oficina Nacional de Lucha Contra los Delitos de Odio (ONDOD) en virtud de la Instrucción núm. 1/2018 de la Secretaría de Estado de Seguridad, que depende de la Dirección General de Coordinación y Estudios del Ministerio del Interior. Hasta el momento, es loable el gran e incesante trabajo realizado por la ONDOD, por medio de una serie de instrumentos accesibles a través de Internet en la web del Ministerio del Interior. A todo ello se le suman dos Planes de Acción de lucha contra los delitos de odio.[3] El primero de ellos, se ejecutó desde 2019 a 2021, y sirvió para otorgar un nuevo enfoque e impulso a la respuesta de las Fuerzas y Cuerpos de Seguridad del Estado (FCSE) ante los incidentes y delitos de odio.

Por ello, dos de los grandes retos procesales vinculados a este tipo de delitos hacen referencia a impulsar una investigación efectiva y una forma de probar eficiente, cuando estos son cometidos en redes sociales abiertas, razón por la que cobran protagonismo los programas para detectar discurso de odio a través de algoritmos. A continuación, partimos de la construcción mediante prueba indiciaria de indicadores de polarización radical para, posteriormente, centrarnos en los principales programas impulsados desde el Ministerio del Interior en la lucha para detectar el discurso de odio en línea y las implicaciones procesales que ello conlleva.

2. Prueba indiciaria y delitos de odio en Internet

La Circular 7/2019 de la Fiscalía General del Estado establece pautas para interpretar los delitos de odio en el artículo 510 CP. Según esta Circular, para atribuir un delito de odio a un acusado es necesario probar tanto el hecho delictivo como la participación del acusado, así como la intencionalidad del autor. Dicho texto enfatiza que el juez debe llegar a esta conclusión al evaluar y analizar las pruebas presentadas, utilizando inferencias o juicios de valor, para tomar una decisión fundamental sobre el caso. En este contexto, un juicio de inferencia implica evaluar pruebas indirectas o circunstanciales para llegar a una conclusión o inferencia sobre un hecho o evento, basándose en elementos probatorios indirectos que permiten inferir la existencia de un hecho base que no se puede probar directamente. En el caso de delitos de odio en Internet, se deben presentar pruebas que demuestren un móvil discriminatorio, es decir, una relación de causa y efecto entre la conducta realizada y la motivación discriminatoria del autor. Para establecer la intencionalidad discriminatoria del autor, se puede recurrir al juicio de inferencia utilizando pruebas indirectas o circunstanciales. En este sentido, la Circular menciona la importancia de utilizar indicadores de polarización como elementos que sugieren la existencia de una actitud discriminatoria y que pueden ser utilizados para

3. *Vid. I Plan de Acción de lucha contra los delitos de odio. Ministerio del Interior*. Gobierno de España. Disponible en: https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/Delitos-de-odio/descargas/PLAN-DE-ACCION-DE-LUCHA-CONTRA-LOS-DELITOS-DE-ODIO.pdf. [Fecha de consulta: 10 de julio de 2023].

inferir la intencionalidad del autor. En resumen, se trata de deducir la intencionalidad discriminatoria del investigado a través de indicios en función del apartado del artículo 510 CP en el que nos encontremos.

Asimismo, la Circular es expresa y clara al afirmar que «la concurrencia de una motivación de odio o discriminación ha de acreditarse normalmente a través de parámetros indiciarios», para posteriormente volver a resaltar con rotundidad que «no se puede desconocer la dificultad que, tradicionalmente, ha existido para valorar la concurrencia de un sentimiento tan íntimo como es la intención que guía al sujeto activo de un hecho delictivo, para lo que debe recurrirse al juicio de inferencia a través de la prueba indiciaria». Al hablar directamente de «prueba indiciaria», debemos conectar con el reconocimiento del derecho a la prueba, vinculado al derecho a la tutela judicial efectiva que todo ciudadano tiene reconocido en el artículo 24 de la Constitución española (CE). Concretamente en el apartado segundo hacemos referencia a que se pueden utilizar todos los medios de prueba necesarios para la defensa que se dispongan en la ley, debido a que las fuentes de prueba son ilimitadas mientras que los medios están tasados en nuestras leyes procesales.

Si bien, concretamente, la prueba indiciaria no encuentra reconocimiento legal expreso como medio de prueba, sino jurisprudencial. Esta falta de reconocimiento legal le acarrea una labor de motivación mucho más dificultosa. Así lo recoge el Tribunal Supremo (TS) de manera concreta, al indicar que la valoración de una prueba directa conlleva una menor complejidad: «es suficiente la indicación de la prueba directa sin que sea preciso, en principio, un especial razonamiento, como, por el contrario, es necesario cuando de pruebas indiciarias se trata».[4]

En el mismo sentido, años después, se pronuncia la STS 947/2007 de 12 de noviembre: «el recelo respecto a la prueba indiciaria no es de ahora. Los aforismos *plus valet quod in veritate est quam quod in opinione o probatio vincit paresumptionem* son la mejor muestra de la preocupación histórica por fijar garantías adicionales que disminuyan el riesgo inherente a la proclamación de unos hechos probados a partir de una mera articulación lógica de indicios».[5] Marchena mantiene esta línea argumental en sentencias posteriores, en las que exige de manera recurrente que estas garantías se den sobre cada uno de los elementos que conforman la prueba de indicios.[6] Concretamente, se exige tanto por el TS como por el Tribunal Constitucional (TC),[7] que los indicios sean plurales, ya que «un solo indicio podría fácilmente inducir a error», además deberán ser necesarios, capaces de sostener válidamente una inferencia presuntiva (de Miranda Vázquez, 2011); en tercer lugar, deberán ser armónicos, entendiendo que deben apuntar en la misma dirección que la línea de defensa elegida; en cuarto lugar, deberán ser referenciados en la sentencia; y, en último lugar, deberán probar el hecho base por vía directa.[8] Esta línea, además, ha sido reforzada por el TS en el 2019 con un nuevo estándar de prueba objetivo aplicable a los indicios en el proceso penal, aludiendo a exigencia de un canon de «probabilidad prevaleciente».[9]

Si nos centramos concretamente en la investigación de delitos de odio en redes sociales, defendemos que la prueba de inteligencia policial partirá de una construcción mediante pruebas de indicios, y que la misma puede ser articulada para probar delitos de odio en Internet. Al respecto, la Circular 7/2019 indica expresamente, justamente después de exponer que los indicadores de polarización deben ser planteados como indicios en estos delitos, que, «en casos de especial gravedad o complejidad puede revestir particular importancia la utilización de la denominada "prueba pericial de inteligencia", cuyo encaje se encuentra en el artículo 370.4 de la Ley de Enjuiciamiento Civil, para adentrarse en las entrañas de determinados colectivos que, precisamente por la opacidad en la que desarrollan sus conductas ilícitas, no permiten contar con otras fuentes de prueba más habituales».

Debemos decir que, acorde con lo expuesto por la Fiscalía, compartimos parcialmente el enfoque con el que se refiere a la prueba de inteligencia. Consideramos que parte de una concepción obsoleta al plantear su uso únicamente

4. STS 126/2003, de 29 de enero, FJ 5.
5. STS 947/2007 de 12 de noviembre, FJ 1.
6. STS 456/2008, de 8 de julio. FJ 1.
7. Estos aspectos son también tratados por el TC en las SSTC31/1981, 160/1988, 150/1989, 109/1986 y 13/1995.
8. Esta línea esta reiterada por el TS desde los años 80 y 90. *Vid*. SSTS de 22 de julio de 1987, de 7 de abril de 1989, de 15 de octubre de 1990, de 13 de mayo de 1996.
9. A nivel doctrinal, analiza se puede comprobar un acertado análisis crítico de la misma en MUÑOZ ARANGUREN, A. (2020, 4 de marzo). «La valoración judicial de la prueba de indicios: una lectura crítica de la STS 532/2019, de 4 de noviembre», *Diario La Ley*, n.º 9586.

en delitos relacionados con organizaciones criminales de manera compleja. Esta visión puede haber sido influida por la jurisprudencia del TS de hace más de una década, la cual se basaba en una configuración jurisprudencial. Sin embargo, creemos que esta perspectiva ha quedado superada debido a las características propias de la investigación en redes abiertas y las técnicas OSINT (del inglés, *Open Source Intelligence*), las cuales permiten obtener indicios de manera más eficiente en este tipo de entornos.

En este sentido, creemos que nos encontramos ante una especie de democratización y acceso universal a la inteligencia de datos, lo que nos lleva a plantear el uso de esta prueba no únicamente para casos complejos, especialmente graves o en los que participen organizaciones criminales, sino que la labor de inteligencia de datos electrónicos puede también plantearse en investigaciones contra particulares que ataquen a personas concretas por su vinculación a un colectivo vulnerable, o a colectivos en su conjunto, en función de las distintas modalidades recogidas en el artículo 510 del Código Penal, y que pueden además derivar en el tratamiento de tipo agravado por medio de la difusión en redes sociales de determinados mensajes constitutivos de delito de odio.

En atención a lo planteado, ¿una prueba de inteligencia construida por indicios debe tener el valor de prueba indiciaria? A nuestro juicio, la respuesta merece una reflexión más compleja, pues se trata de dos cuestiones diferentes. La generación de esta prueba implica la combinación de varios indicios y, al mismo tiempo, la aplicación de una labor de inteligencia realizada por el agente investigador. Por tanto, podemos decir que esta prueba de inteligencia no está únicamente construida por indicios, sino también por la pericia de profesionales que aportan conocimiento especializado, tanto en la técnica de investigación para recabar dichos indicios como en la manera de entrelazarlos e interpretarlos.

Para sostener de manera adicional este planteamiento, concluimos compartiendo una postura relativamente abrupta defendida en los últimos pronunciamientos del TS, en la que habla de prueba indiciaria como una forma de razonar los hechos base pero no como un nuevo medio de prueba, al indicar, respecto a la prueba indiciaria, que, «más que una prueba estaríamos en presencia de un sistema o mecanismo intelectual para la fijación de los hechos, ciertamente relacionado con la prueba, pero que no se configura propiamente como un medio de prueba». De igual modo, Gimeno Sendra (2015) indicaba que «la prueba por indicios forma parte del juicio de hecho, pero no como un medio de prueba que es valorado, sino como una operación intelectual o técnica de prueba, por lo que es propio de la fase de valoración de prueba», o Gómez Colomer (2014), quien indica que más bien sería «un método de prueba de aplicación general a cualquier tipo de delitos (...) no es más que un esquema de razonamiento que cabe utilizar a propósito de cualquier medio de prueba».

Por tanto, compartiendo la postura de estos autores, llegamos a la conclusión de que los indicios no pueden ser considerados como medio de prueba, sino más bien como elementos que configuran una forma concreta de razonamiento probatorio y, a su vez, si los medios son los que se establecen en nuestras leyes procesales, tendremos que ver en qué medio de prueba concreto se concreta la prueba de inteligencia policial. Pero antes, en términos más específicos, es necesario examinar qué indicios concretos deben considerarse para construir dicha prueba. En el contexto de los delitos de odio, estos indicios son comúnmente conocidos como «indicadores de polarización radical», los cuales se analizarán a continuación.

3. Indicadores indiciarios que orientan la investigación y la prueba de los delitos de odio. Especial referencia al uso de algoritmos[10]

Tal y como se recoge en la guía de referencia titulada *Preventing and responding to hate crimes*,[11] publicada por la OSCE y la Oficina para las Instituciones Democráticas y los Derechos Humanos (ODIHR) en 2009, se indi-

10. Igualmente, sobre OSINT, BREZO FERNÁNDEZ, F.; RUBIO VIÑUELA, Y. (2019). *Manual de ciberinvestigación en fuentes abiertas. OSINT para analistas*. Madrid, pág. 3. RODRÍGUEZ RODRÍGUEZ, Y. (2019). «Inteligencia de fuentes abiertas (OSINT): Características, debilidades y engaño». *Revista de Análisis* GESI, n.º 11; ORTEGA, J. M. (2021). *Herramientas OSINT para auditorías de seguridad y ciberamenazas. Obteniendo inteligencia a partir de fuentes abiertas*. Madrid, pág. 3;
11. Disponible en: https://www.osce.org/files/f/documents/8/a/39821.pdf. [Fecha de consulta: 20 de marzo de 2023].

ca que, «a la hora de investigar un delito de odio, el problema más corriente es la negativa o la incapacidad de las autoridades para identificar un acto criminal como un delito de odio. Por ello, es esencial que los agentes de policía y los representantes de las ONG que reciben las denuncias o entrevistan a las víctimas dispongan de criterios que les permitan determinar si se trata de un delito de odio». Es decir, una de las principales dificultades a las que se enfrentan los FCSE es ver que estamos ante un verdadero delito de odio, debido a la dificultad para determinar la motivación que subyace detrás del sujeto investigado.

Desde la perspectiva procesal, el problema se concreta en saber a través de qué mecanismos podemos identificarlos, y es ahí cuando por medio de la Sentencia del Tribuanl Europeo de Derechos Humanos (TEDH) del 20 de octubre de 2015. Caso «Balázs vs. Hungría» (*Application* n.º 15529/12) se plantean los denominados «indicadores de polarización radical». Estos indicadores nos proporcionan orientación sobre cómo investigar un hecho en particular como un delito de odio, al mostrar que, tanto la intención como la motivación pueden deducirse de ellos. En este sentido, estamos nuevamente frente a un razonamiento inferencial y, por lo tanto, claramente ante una prueba indiciaria. Es decir, estaríamos hablando de circunstancias o acciones del agresor que, cuando se consideran de forma individual o en conjunto con otros factores, sugieren que el delito fue motivado por odio o discriminación hacia una persona o grupo específico.

No obstante, la presencia de uno o varios de estos indicios no constituye una prueba concluyente de que se haya cometido un delito de odio, sino más bien una evidencia circunstancial de la motivación subyacente. Por lo tanto, es esencial que los investigadores y los tribunales analicen minuciosamente cada caso para determinar si se han satisfecho los requisitos necesarios para clasificar un delito como un crimen de odio.

Es importante destacar que, aunque los indicadores de polarización pueden ser una herramienta útil para establecer la motivación detrás de un delito de odio, no son suficientes por sí solos para probar la culpabilidad del acusado. En su lugar, los investigadores y tribunales deben considerar estos indicadores en conjunto con otras pruebas, tales como testimonios de testigos presenciales, pruebas periciales, y cualquier otra evidencia pertinente.

Por lo tanto, los indicadores de polarización pueden servir como una guía útil para la investigación y el tratamiento policial de un delito de odio, siempre y cuando existan múltiples indicadores que converjan. Sin embargo, es fundamental tener en cuenta que estos indicadores no son suficientes por sí solos para demostrar la culpabilidad del acusado y deben ser evaluados junto con otras pruebas para determinar si se han cumplido los requisitos legales para considerar un delito como un delito de odio.

Ahora bien, la numeración de indicadores de polarización no es homogénea en función del protocolo que se siga, pues se parte de una interpretación de la STS del TEDH aludida. Estos mismos criterios han sido además recogidos a nivel de organizaciones internacionales tanto por la OSCE y la Comisión Europea contra el Racismo y la Intolerancia (ECRI), esta última recogiendo a su vez los criterios del Plan de Acción de Rabat de Naciones Unidas para fijar el umbral que permita establecer adecuadamente qué tipo de expresiones pueden constituir delito de odio. Concretamente, la recomendación n.º 15 de la ECRI establece criterios para determinar si ciertas expresiones constituyen delito de odio, considerando el contexto, la influencia del emisor, el lenguaje utilizado, la naturaleza de los comentarios, el medio utilizado y la audiencia involucrada. Estos criterios ayudan a evaluar si se incita a la violencia, la intimidación, la hostilidad o la discriminación. En el ámbito nacional, la Circular de la Fiscalía General del Estado (CFGE) del 2019 organiza estos indicadores de polarización reduciéndolos a tres bloques (autor, víctima y contexto), dejando a alguno de ellos fuera porque suponemos entender que no encajan en ninguna de las tres categorías.

El traslado de estos indicadores de polarización radical a las redes sociales representa una cuestión más compleja. A falta de consecución de un proyecto europeo, que posteriormente mencionaremos, a fecha de redacción del presente artículo no existe un traslado de los indicadores de polarización específicos, ni un protocolo concreto a este respecto. Si bien contamos con un *Protocolo para combatir el discurso de odio ilegal en línea (#ContraeldiscursodeOdio)*, suscrito en febrero de 2021, el cual se configura como un instrumento para la colaboración efectiva entre los actores que se ocupan de la lucha contra el discurso de odio ilegal en línea en España y en diferentes países de la UE: instituciones de la Ad-

ministración Pública, organizaciones de la sociedad civil y prestadores de servicios de alojamiento de datos.[12] El protocolo busca fomentar la cooperación y coordinación entre autoridades nacionales y europeas para agilizar la investigación de delitos de discurso de odio, estableciendo objetivos como, definir los delitos, elaborar un listado de autoridades competentes y promover mecanismos de resolución extrajudiciales. Además, propone la tramitación preferente de comunicaciones de comunicantes fiables, la creación de un sello de acreditación y formación, y la creación de una comisión de seguimiento para garantizar el cumplimiento del texto.

Si bien, aunque se asume que el discurso de odio se encuentra tipificado en el artículo 510 CP, también podría quedar incluido entre los delitos tipificados en la legislación penal española consistentes en actos expresivo-comunicativos a los que fuera de aplicación el artículo 22.4ª del CP, así como en los apartados b) y c) del artículo 23.1 de la Ley 19/2007 de 11 de julio, contra la violencia, el racismo, la xenofobia y la intolerancia en el deporte; siempre y cuando se trate de conductas desarrolladas en la red que estén alojadas en servidores y, a su vez, que respete la normativa europea al respecto. No obstante, no se concreta ni propone la creación de un listado de indicadores de polarización concreto a la hora de investigar las redes sociales abiertas del autor, con el fin de motivar y probar el discurso de odio en Internet.

De igual modo, en el *Protocolo de actuación de las FCSE para los delitos de odio y conductas que vulneran las normas legales sobre discriminación* se indica que para los agentes «se plantea una disyuntiva difícil de dirimir en algunos supuestos, debido a que los miembros de las Fuerzas y Cuerpos de Seguridad deben determinar si los contenidos difundidos mediante los medios de comunicación electrónicos constituyen un ataque directo a una persona o a un colectivo especialmente vulnerable o por el contrario constituye un ejercicio de la libertad de expresión», prestando una especial atención al análisis de música compartida en plataformas digitales y procurando el equilibrio entre libertad de expresión y odio y, posteriormente, se reconduce para su investigación y detección a lo dispuesto en lo establecido en la CFGE 7/2019 y en la Recomendación n.º 15 de la ECRI, que recoge los criterios del Plan de Acción de Rabat de Naciones Unidas redirigiéndonos a los indicadores de polarización apuntados previamente de manera genérica.

Si acudimos al II Plan de Acción Nacional de Lucha Contra los Delitos de Odio, vemos como se cuenta con una iniciativa vinculada a promover el protocolo anterior a través de convertir a la ONDOD en una figura de *trusted flagger* (comunicante fiable), participando en el ejercicio anual de monitoreo de los proveedores de servicios de Internet, con base en el *Código de Conducta para contrarrestar el discurso de odio en línea* y en el que la ONDOD se compromete a promover la realización de contactos periódicos con los distintos prestadores de servicios de Internet. Podemos, además, afirmar que esta función seria conforme la propuesta de Directiva del Parlamento Europeo y del Consejo, por la que se establecen normas armonizadas para la designación de representantes legales a afectos de recabar pruebas de naturaleza electrónica en procesos penales, y de manera más concreta en su artículo 3 (Bujosa Vadell, 2022, pág. 97).

4. Algoritmos e identificación de indicadores de polarización radical del discurso de odio en línea

¿Se puede detectar de manera automatizada el delito de odio a través de la identificación de indicadores de polarización en redes sociales? En este caso, la inteligencia artifical (IA) hace su irrupción.

Los programas informáticos que ayudan a detectar los delitos de odio en general, y el discurso de odio en línea en particular, son esenciales para combatir este problema creciente, puesto que proporcionan una forma eficaz de identificar el contenido ofensivo, y ayudan a prevenir su propagación. Por este motivo, es importante que no solo las empresas o las iniciativas privadas, sino también que las Administraciones Públicas, sigan invirtiendo en la investigación y el desarrollo de estas tecnologías para garantizar que las personas que pertenecen a los colectivos vulnerables no sufran los efectos negativos de los delitos de odio en Internet, ya que, al fin y al cabo, nos encontramos ante acciones delictivas que por definición tienen

12. *Vid. Protocolo para combatir el discurso de odio ilegal en línea (#ContraeldiscursodeOdio)*. Ministerio de Justicia. Febrero de 2021. Disponible en: https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/Delitos-de-odio/descargas/protocolo-discurso-odio.pdf. [Fecha de consulta: 20 de febrero de 2023].

interés público. Por esta razón, creemos necesario hacer alusión a los programas informáticos impulsados a nivel institucional y gubernamental y, más concretamente, a aquellos que han tenido relación en su origen y desarrollo con la ONDOD, debido a la particular situación que esta institución representa en nuestro país

En este sentido, la ONDOD no existe como una entidad oficial en todos los países más allá de España, por lo que no es fácil realizar una equiparación de esta institución a nivel de derecho comparado. En cuanto a herramientas, y a pesar de que la ONDOD no pueda indicar qué *softwares* específicos utiliza de manera taxativa para detectar delitos de odio en redes sociales por razones obvias vinculadas al secreto profesional y al propio de cualquier fase de instrucción, es probable que utilice una combinación de herramientas y técnicas para detectar y analizar contenido ofensivo de contenido múltiple y, al mismo tiempo, ayude, desde su sección de estudios, a construir programas que permitan detectar el discurso en redes sociales y publiquen los resultados de investigación pertinentes. Es decir, realmente hablamos de utilización de programas SOCMINT (del inglés, *Social Media Intelligence*), pero que a su vez aglutinen técnicas de OSINT, SIGINT (del inglés, *Signals Intelligence*) y IMINT (del inglés, *Imagery Intelligence*), puesto que las funcionalidades de esas herramientas están vinculadas a este tipo de inteligencia.

Concretamente, la ONDOD impulsa una propuesta de algoritmo para detectar discurso de odio, HaterNet en su primera versión y posteriormente SocialHaterbert, como continuación de la primera, presentadas a través de artículos científicos en revistas de impacto. De igual modo, la ONDOD participó en un proyecto europeo, denominado ALRECO, que propuso los estándares para generar algoritmos para la detección de delito de odio y participa en otro que, a fecha de redacción del presente estudio, se encuentra en activo, denominado Social Real-UP. En el año 2019 se publica un artículo presentando HaterNet (Pereira-Kohatsu; Quijano-Sánchez; Liberatore; Camacho-Collados, 2019), un sistema inteligente impulsado por la ONDOD para detectar de manera automática el discurso de odio en línea, con especial atención a la red social Twitter. Posteriormente, en diciembre de 2022, se publica un segundo artículo para presentar la evolución del programa HaterNet, ahora denominado SocialHaterBERT (Valle-Cano; Quijano-Sánchez; Liberatore; Gómez Esteban, 2023), el cual vuelve a estar centrado en identificar y monitorear la evolución del discurso de odio en Twitter, pero mejorando el algoritmo creado en 2019 a través de la incorporación de la tecnología BERT (*Bidirectional Encoder Representations from Transformers*). BERT es un modelo de lenguaje desarrollado por Google en 2018. Es una red neuronal basada en la arquitectura de transformadores (Scola; Segura Bedmar, 2021), que permite entender el contexto en el que se usan las palabras en un texto a través de técnicas de programación vinculadas al aprendizaje automático y al procesamiento del lenguaje natural, como la comprensión de texto, la generación de texto, la traducción automática y la clasificación de texto, con un alto nivel de precisión. Es ampliamente utilizado en aplicaciones de IA, y se considera uno de los modelos de lenguaje más avanzados actualmente disponibles mundialmente.

En cuanto al discurso de odio, lo que se pretende es que BERT sea utilizado para detectar el discurso de odio en línea mediante el entrenamiento del modelo de HaterNet, con datos etiquetados como discurso de odio o no discurso de odio en la base de *tweets* que hemos referenciado en el apartado anterior (6000 *tweets* etiquetados, descargados entre febrero y diciembre de 2017) y, a partir de ese modelo, clasificar los mensajes. Es decir, este modelo lo que hace es favorecer una compresión contextual del *tweet*, y no solo se queda en palabras determinadas, por lo que, según los autores, nos encontraríamos ante el mejor modelo para clasificar discurso de odio en español. Así, este algoritmo base, HaterBERT, mejora los resultados de los clasificadores españoles en un 3 % a 27 %.

De igual modo, se potencia una nueva metodología para el análisis de *tweets* denominada SocialGraph, consistente en una base de datos de usuarios de Twitter que incluye características textuales y numéricas de los perfiles de usuario, actividad pasada y entorno. Es decir, serían una serie de medidas de inteligencia social que definen características de los usuarios en la Red y cómo se comportan en un entorno virtual determinado.

Por último, se plantea integrar las dos medidas anteriores en un modelo final de algoritmo rebautizado como SocialHaterBERT. Es decir, sería un modelo algoritmo que va más allá del análisis de texto, al que se suma la interpretación de características y comportamiento del usuario que lo definen socialmente dentro de esa red. En definitiva, lo que se ofrece es un nuevo modelo, innovador y superior al anterior, que combina datos sociales y texto para mejorar la identificación del discurso de odio en redes sociales.

Una vez planteados estos dos algoritmos, ¿cómo podría encajar su uso en relación con los indicadores de polarización anteriormente planteados? Es aquí cuando cobra

sentido el proyecto europeo Discurso de odio, racismo y xenofobia: mecanismos de alerta y respuesta coordinada, bautizado con el acrónimo ALRECO, vigente de noviembre de 2018 a abril de 2021.[13] Su meta era mejorar la capacidad de las autoridades del Estado para detectar, analizar, monitorear y evaluar el discurso de odio en las redes sociales por medio de la identificación de factores de polarización, con el objetivo de crear estrategias conjuntas contra el discurso motivado por el racismo, la xenofobia, la islamofobia, el antisemitismo y el antigitanismo.

Lo fundamental de este proyecto es que identifica los indicadores que se usarán para detectar el discurso de odio en las redes sociales a través de un algoritmo modelo que pretenden crear. En este sentido, en la herramienta de monitoreo se utilizarán diferentes tipos de indicadores para determinar si un *tweet* es considerado discurso de odio o no, y para clasificar su intensidad. Algunos de estos indicadores incluyen la presencia de palabras específicas, el lenguaje que incita a la violencia, la justificación de la violencia, la reproducción de estereotipos, etc., y los trata de diferenciar de las meras bromas, prejuicios, rumores, datos fasos, falacias, argumentos trampas, lenguaje malsonante o despectivo.

Una vez el algoritmo sepa separar lo que constituye odio de lo que no, hará una clasificación de cada tipo de discurso. Así, si el test es positivo y se indica que hay componente de odio, planteará si el discurso de odio tiene un grado extremo, que supone incitar a la violencia contra un colectivo o persona por pertenecer a este; u ofensivo, que incita a la discriminación, reproduciendo tópicos u estereotipos contra colectivos vulnerables. En cambio, si el algoritmo indica que no existe componente de odio, logrará diferenciar si nos encontramos ante un discurso nuestro basado en un enfoque descriptivo, o si estamos ante un discurso alternativo o contranarrativo, es decir, que cuente una realidad del odio que manifiestan terceros, o que use una expresión de odio utilizada por otros para refutarlo, no siendo el investigado autor, sino defensor de los derechos de un determinado colectivo o que simplemente manifieste una opinión distinta sin intención discriminatoria.

En este sentido, nos encontramos ante un algoritmo que utilizaría técnicas de aprendizaje automático para detectar discurso de odio en redes sociales a través de un primer análisis cuantitativo, mediante la selección de un banco de palabras determinadas y un conjunto de *tweets*, elegidos por determinados parámetros y, posteriormente, por medio de un análisis cualitativo, enfocado en clasificar el contenido de los *tweets* según la intensidad de odio a través de técnicas de *machine learning* supervisado. El proceso incluye la limpieza de los datos y el etiquetaje manual de los *tweets* por personas revisoras. La finalidad es identificar patrones en el lenguaje y homogeneizar el contenido para enfocarse en la intencionalidad propia de cada *tweet*, para lo que será muy importarte entrenar al algoritmo. Una vez entrenado, se crea un modelo a partir de los *tweets* etiquetados manualmente, que será capaz de clasificarlos según la intensidad de odio. La clasificación o predicción se realizará aplicando el modelo a los *tweets* no clasificados, mediante un procedimiento reiterativo y periódico, que se realizará actualizando nuevos conjuntos de *tweets* clasificados de manera progresiva.

Todo ello, entronca con una serie de medidas que se recogen en el II Plan de Acción contra los Delitos de Odio, que deben ser desarrollados e implementados a nivel legal antes de finalizar el 2024. Así, untos 7.2 y 7.3 de la Estrategia, a desarrollarse en el segundo semestre de 2024:

• «7.2. Impulsar las reformas normativas o legislativas necesarias al objeto de avanzar en la lucha contra el discurso de odio en línea y los delitos de odio en general, principalmente en el ámbito administrativo.

• 7.3. Desarrollo de un análisis espacio-temporal de los delitos de odio y su relación con el discurso de odio, al objeto de conocer si existe alguna relación/correlación entre el discurso de odio en línea y los delitos de odio en el "mundo físico" o, viceversa».[14]

En un contexto actual, debemos citar el Proyecto Europeo Real-UP, el cual comenzó en diciembre de 2021 y tiene vigencia hasta diciembre del presente 2023.[15] Este

13. *Action Grant* (AL-RE-CO) con referencia «Just/2017/Action Grants /REC PROGRAM» [en línea]. Disponible en: https://www.inclusion.gob.es/oberaxe/es/ejes/delitosodio/alreco/index.htm
14. *II Plan de acción de lucha contra los delitos de odio. Ministerio del Interior*. Secretaría de Estado de Seguridad. Abril 2022. *Vid*. https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/Delitos-de-odio/descargas/II-PLAN-DE-ACCION-DE-LUCHA-CONTRA-LOS-DELITOS-DE-ODIO.pdf. [Fecha de consulta: 11 de marzo de 2023].
15. Proyecto europeo REAL – UP. Disponible en: https://real-up.eu/. [Fecha de consulta: 5 de febrero de 2023].

comparte una base similar con el proyecto ALRECO en términos de perfilación de la herramienta algorítmica y se encuentra liderado por el Observatorio Español del racismo y la Xenofobia (OBERAXE). El proyecto tiene como objetivo principal mejorar la capacidad de las autoridades estatales para detectar, evaluar y supervisar el discurso de odio en línea y desarrollar estrategias efectivas de contranarrativa o generación de discurso positivo contra el discurso de odio motivado por el racismo, la xenofobia, la islamofobia, el antisemitismo y el antigitanismo.

Todo ello conecta con analizar la efectividad del discurso *upstander* como una herramienta contra el discurso de odio en línea, y dicha narrativa también se podrá identificar por medio de herramientas de inteligencia SOCMINT, mediante el análisis del perfil del autor y la víctima del discurso de odio mediante algoritmos centrados en modelo BERT y con aplicación de técnicas de *web scraping* y *crawling*.

5. A modo de conclusión

Se presentan dos desafíos en la prueba de los delitos de odio en las redes sociales: probar la intención y la motivación. La prueba de la motivación es más difícil que la de la intención, ya que se refiere a un aspecto interno de la mente del autor de los hechos. Ambos desafíos requieren considerar ciertos criterios y delimitar el odio en comparación con otros delitos y la libertad de expresión.

Se han desarrollado «indicadores de polarización radical» para sugerir la posible comisión de un delito de odio y probar así la intención y la motivación detrás del delito. Estos indicadores son considerados como prueba indiciaria, una institución que actualmente no cuenta con reconocimiento legal expreso, y si atendemos a su construcción jurisprudencial no existe unanimidad en considerarla como un medio de prueba autónomo, ni tampoco en cuanto a sus características y tipología.

Pero ¿cómo podríamos investigar de una manera eficiente este tipo de delitos cuando están asociados con el discurso de odio en línea? Sin duda a través de un modelo de inteligencia de fuentes. Dentro del modelo de inteligencia de fuentes, la técnica o metodología OSINT será la idónea para investigar delitos de odio en redes sociales, puesto que la interrelación de datos electrónicos masivos ofrecidos por el autor será la principal fuente de prueba en la que se basará. De manera complementaria, se podrán aplicar técnicas secundarias como IMINT y SIGINT, por lo que nos encontraríamos ante un tipo de «inteligencia de fuentes holística» (Bueno de Mata, 2023).

OSINT ejemplifica un modelo de inteligencia completamente distinto respecto a la concepción clásica de inteligencia de hace tres décadas, adaptado a la realidad tecnológica e hiperconectada en la que vivimos hoy en día, y en la que los datos abiertos marcan una nueva línea para investigar determinadas situaciones y, más concretamente, hechos delictivos cometidos por particulares. Por todo ello, defendemos que OSINT desafía por sí mismo el concepto clásico de inteligencia, ya que, conforme a dicha conceptualización, la inteligencia debería actuar sin publicidad ni trasparencia, sin que sus objetivos y sus métodos fuesen públicos.

En este sentido, las técnicas OSINT irán ganando protagonismo de manera clara en el futuro a la hora de investigar el discurso de odio en Internet, basándose en indicadores de polarización que puedan llevarse al proceso por medio de una prueba de inteligencia policial que sirva para interpretar de manera unida y correlacionada una serie de fuentes de prueba indiciarias, siempre que la legislación no las limite.

En este sentido, la versión consolidada, a fecha de julio de 2023, del Reglamento para la regulación de la IA, establece la prohibición de los sistemas de policía predictiva y reincidencia criminal, por lo que se tendrá que ver de qué manera casan estas limitaciones con los algoritmos planteados (Estévez Mendoza, 2019) y, en tal caso, si España se acoge al impulso de un *sandbox* regulatorio para dirimir esta cuestión, razón que deberá ser debatida antes de que el II Plan de Acción contra los Delitos de Odio finalice y se tenga una respuesta clara y perfilada, nunca mejor dicho, jurídicamente, antes de finalizar el año 2024.

Reconocimientos

Esta publicación forma parte del proyecto nacional de I+D+i Tratamiento procesal de los delitos de odio cometidos a través de medios tecnológicos, ref. PID2021-128339OA-I00, perteneciente a la convocatoria sobre Proyectos de generación de conocimiento del Plan Estatal de Investigación Científica, Técnica y de Innovación 2021-2023; financiado por MCIN/ AEI /10.13039/501100011033/ y por FEDER: Una manera de hacer Europa. IP. BUENO DE MATA.F.

Referencias bibliográficas

BUENO DE MATA, F. (2023). *Investigación y prueba de delitos de odio en redes sociales: técnicas OSINT e inteligencia policial*. València: Tirant lo Blanch.

BREZO FERNÁNDEZ, F.; RUBIO VIÑUELA, Y. (2019). *Manual de ciberinvestigación en fuentes abiertas. OSINT para analistas*. Madrid.

BUJOSA VADELL, L. (2022). «Cooperación judicial para la obtención y transmisión de pruebas electrónicas». *A vueltas con la transformación digital de la cooperación jurídico penal internacional*, págs. 97-123. Navarra: Marcial Pons.

DE MIRANDA VÁZQUEZ, C. (2011, enero). «Indicios y presunciones en la doctrina jurisprudencial de la Sala 2º del Tribunal Supremo». *Diario La Ley*, n.º 7549, págs. 1-11.

DEL VALLE-CANO, G.; QUIJANO-SÁNCHEZ, L.; LIBERATORE, F.; GÓMEZ ESTEBAN, J. (2023, abril). «SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles». *Expert Systems with Applications*, vol. 216. 119446, pags.1-17. DOI: https://doi.org/10.1016/j.eswa.2022.119446

ESTÉVEZ MENDOZA, L. (2019). «Algoritmos policiales basados en IA y derechos fundamentales a la luz de HART y VALCRI: garantías versus eficacia». *Justicia: ¿garantías versus eficacia?*, págs. 665-674. València: Tirant lo Blanc.

GIMENO SENDRA, V. (2015). *Derecho Procesal Penal*, pág. 248. Pamplona: Civitas.

GÓMEZ COLOMER, J.M. (2014). *Derecho Jurisdiccional*. Tomo III, pág. 211 y ss. Valencia: Tirant lo Blanch.

MARTÍN RIOS, P. (2022). *Digital Forensics and criminal process in Spain: evidence gathering in a changing context*. Navarra: Aranzadi.

MIRÓ LLINARES, F. (2017). «Derecho penal y 140 caracteres. Hacia una exégesis restrictiva de los delitos de expresión». *Cometer delitos en 140 caracteres. El Derecho penal ante el odio y la radicalización en Internet*. Madrid: Marcial Pons.

MUÑOZ ARANGUREN, A. (2020, marzo). «La valoración judicial de la prueba de indicios: una lectura crítica de la STS 532/2019, de 4 de noviembre». *Diario La Ley*, n.º 9586.

ORTEGA, J.M. (2021). *Herramientas OSINT para auditorías de seguridad y ciberamenazas. Obteniendo inteligencia a partir de fuentes abiertas*, pág. 3. Madrid.

PEREIRA-KOHATSU, J.C.; QUIJANO-SÁNCHEZ, L.; LIBERATORE, F.; CAMACHO-COLLADOS, M. (2019). «Detecting and Monitoring Hate Speech in Twitter». *Sensors*, vol. 19, n.º 21, págs. 1-37. DOI: https://doi.org/10.3390/s19214654

RODRIGUEZ RODRÍGUEZ, Y. (2019). «Inteligencia de fuentes abiertas (OSINT): Características, debilidades y engaño». *Revista de Análisis GESI*, n.º 11.

SCOLA, E.; SEGURA BEDMAR, I. (2021). «Detección de Sarcasmo con BERT». *Procesamiento del lenguaje natural*, n. º 67. págs. 13-25. DOI: https://doi.org/10.26342/2021-67-1

VALLE-CANO, G. DEL; QUIJANO-SÁNCHEZ, L.; LIBERATORE, F.; GÓMEZ ESTEBAN, J. (2023, abril). «SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles». *Expert Systems with Applications*, vol. 216, págs.1-17. DOI: https://doi.org/10.1016/j.eswa.2022.119446.

Cita recomendada

BUENO DE MATA, Federico (2023). «Particularidades probatorias del discurso de odio en Internet: identificación de indicadores de polarización radical mediante sistemas algorítmicos». En: Miró, F. (coord.). «Digitalización y algoritmización de la justicia». *IDP. Revista de Internet, Derecho y Política*, n.º 39. UOC [Fecha de consulta: dd/mm/aa]
http://dx.doi.org/10.7238/idp.v0i39.416359

Sobre las autorías

Federico Bueno de Mata
Universidad de Salamanca
febuma@usal.es

Catedrático de Derecho Procesal de la Universidad de Salamanca (USAL). Actualmente vicedecano de la Facultad de Derecho. Doctor en Derecho con Premio Extraordinario de Doctorado por la USAL con su tesis sobre prueba electrónica. Imparte docencia en Derecho Procesal en distintos grados y posgrados en la USAL, e igualmente es docente invitado en másteres y doctorados en distintas universidades nacionales e internacionales. Sus líneas de investigación están centradas en el derecho procesal y las nuevas tecnologías, la mediación, la tutela judicial de la violencia de género, los menores infractores o las diligencias de investigación tecnológica. Tiene reconocidos 3 sexenios por la Comisión Nacional Evaluadora de la Actividad Investigadora (CNEAI), dos de investigación y otro de transferencia. Autor de seis monografías y más de 130 publicaciones científicas, entre las que se incluyen artículos en prestigiosas revistas jurídicas y contribuciones en libros colectivos sobre temas como el derecho y las nuevas tecnologías, cuestiones probatorias, igualdad, mediación, investigación criminal, etc. Además, es director de la colección de estudios sobre derecho y nuevas tecnologías, FODERTICS. También es ponente en congresos, jornadas y encuentros científicos, tanto en el ámbito nacional como internacional.