Revista Latina de Comunicación Social
ISSN 1138-5820 / No. 82, 01-18.
https://doi.org/10.4185/rlcs-2024-2205

Aplicación de herramientas de IA como metodología para el análisis de la toxicidad en redes sociales: Estudio de caso de la política española en Twitter

Application of AI tools as methodology for the analysis of toxicity in social media: A case study of Spanish politics on Twitter

Uxía Carral
Universidad Carlos III de Madrid. España.
ucarral@hum.uc3m.es

Carlos Elías
Universidad Carlos III de Madrid. España.
carlos.elias@uc3m.es

Financiación: El presente texto nace en el marco del dos proyectos concedidos a miembros de la Universidad Carlos III de Madrid de los cuales recibe financiamiento esta investigación: (1) 'Pseudociencia, teorías conspirativas, fake news y alfabetismo mediático en la comunicación en salud' (Ref: PID2022-142755OB-I00) dentro del Plan Nacional 'Proyectos de Generación de Conocimiento' del Ministerio de Ciencia e Innovación y (2) 'UE, desinformación y fake news' (Ref: 610538-EPP-1-2019-1-ES-EPPJMO-CHAIR), una Cátedra Jean Monnet de la Comisión Europea.

Agradecimientos: Los autores desean agradecer sinceramente el valioso asesoramiento a Rubén Míguez Pérez, Doctor en Ingeniería de Telecomunicaciones por la Universidad de Vigo, que hizo también posible la realización de esta investigación.

Cómo citar este artículo / Referencia normalizada:

Carral, Uxía y Elías, Carlos (2024). Aplicación de herramientas de IA como metodología para el análisis de la toxicidad en redes sociales: Estudio de caso de la política española en Twitter [Application of AI tools as methodology for the analysis of toxicity in social media: A case study of Spanish politics on Twitter]. *Revista Latina de Comunicación Social*, 82, 01-18. https://www.doi.org/10.4185/RLCS-2024-2205

Fecha de Recepción: 04/08/2023
Fecha de Aceptación: 26/12/2023
Fecha de Publicación: 24/01/2023

RESUMEN

Introducción: Se analiza una nueva metodología de inteligencia artificial (IA), entendiendo que la comunicación se presenta como uno de los campos de trabajo más trascendentes para su aplicación. Además de las fases de

recolección y producción de contenido, otras áreas dentro del mundo de la comunicación como la distribución, y en concreto la moderación de comentarios (en redes sociales y en medios) también están viviendo un período de innovación, pero de forma menos evidente para la audiencia. **Metodología:** Se procede a conocer cómo diversas herramientas de IA pueden medir la calidad de la conversación y combatir la toxicidad en espacios comunicativos. Se han analizado 43.165 tuits publicados del 18 al 24 de octubre de 2021 correspondientes a siete políticos españoles y a la cascada de respuestas de los usuarios. **Resultados:** Las principales consecuencias apuntan a los insultos como la categoría tóxica predominante en los comentarios, independientemente de la ideología. Además, las conversaciones cuentan con un promedio del 21% de usuarios *bots*. **Discusión:** Visto lo anterior, esta investigación muestra cómo nuevas metodologías de IA pueden contabilizar un término hasta ahora tan cualitativo como la toxicidad y contradice los hallazgos previos sobre *bots* como difusores de toxicidad, siendo los usuarios reales quienes más toxicidad generan. **Conclusiones:** En el estudio concreto de política, se percibe una diferencia de comportamientos entre la conversación horizontal entre pares y la vertical con los políticos. Por tanto, estas herramientas ayudan a visibilizar nuevas realidades como la toxicidad, con el fin último de llegar a erradicarla y sanear el debate online.

Palabras clave: Inteligencia artificial; Toxicidad; *Bots*; Redes sociales; Twitter; Metodología; Perspective API.

ABSTRACT

Introduction: A new artificial intelligence (AI) methodology is analyzed, with the understanding that communication is one of the most important fields of work for its application. In addition to the content collection and production phases, other areas within the world of communication such as distribution, and specifically the moderation of comments (on social networks and in the media) are also experiencing a period of innovation, but in a less obvious way for the audience. **Methodology:** We proceeded to find out how various AI tools can measure the quality of the conversation and combat toxicity in communicative spaces. We analyzed 43,165 tweets published from 18 to 24 October 2021 corresponding to seven Spanish politicians and the cascade of user responses. **Results:** The main consequences point to insults as the predominant toxic category in the comments, regardless of ideology. In addition, the conversations have an average of 21% of bots. **Discussion:** Given the above, this research shows how new AI methodologies can account for a hitherto qualitative term such as toxicity and contradicts previous findings on bots as spreaders of toxicity, with real users generating the most toxicity. **Conclusions:** In the specific study of politics, there is a perceived difference in behaviors between horizontal conversation between peers and vertical conversation with politicians. Therefore, these tools help to make visible new realities such as toxicity, with the ultimate aim of eradicating it and cleaning up the online debate.

Keywords: Artificial intelligence; Toxicity; Bots; Social media; Twitter; Methodology; Perspective API.

1. Introducción

En 1997, Goldhaber teorizaba sobre la "economía de la atención" aplicada a los procesos de comunicación digitalizados. Apuntaba a la existencia de una esfera virtual como detonante del crecimiento de la monitorización de datos hasta el punto de que la humanidad no sería capaz de abarcar tal cantidad de información (Goldhaber, 1997). Ese fenómeno Toffler (1975) ya lo acuñó en la década de los años setenta como *information overload*, es decir, grandes cantidades de información difícilmente asumibles que, en consecuencia, provocan que la atención haya pasado a ser ese bien escaso, limitado y preciado en detrimento de la información.

Este nuevo escenario mediático centrado en captar la atención de las audiencias ha consolidado un ecosistema proclive a la generación de toxicidad (Southwell *et al*., 2018; Blanco-Alfonso *et al*., 2019). Sin embargo, Silverman (2014) abogaba por el uso de la tecnología como instrumento en la búsqueda de la verdadera información. Esa misma concepción comparte el foco de este trabajo de investigación: determinar el nivel de toxicidad y la naturaleza de la audiencia en la conversación *online* a través de una nueva metodología experimental con diversas herramientas de inteligencia artificial (IA).

"Con una probabilidad CAP otorgada por el propio algoritmo Botometer del 80%"

Para ello, se ha acotado el tema al estudio de la conversación política española en Twitter, no obstante, se podrían escoger otros argumentos discursivos en redes como los terraplanistas o los antivacunas, e incluso, contextos temporalmente concretos como el asalto al Capitolio en Estados Unidos. En definitiva, este estudio de caso es un mero ejemplo que sirve para testear los instrumentos que se proponen, a modo exploratorio, teniendo en cuenta sus posibles limitaciones y errores, pero también abriendo la puerta a la emergencia de una nueva técnica metodológica para el estudio de las redes sociales.

1.1. El origen de la toxicidad en redes sociales

Aunque el concepto de *fake news* se popularizó durante la elección de Trump como presidente de los EE.UU., Wardle y Derakhshan (2017) insisten en ampliar el espectro de 'desinformación' también a aquel contenido con contexto falso, maleducado, hiriente, manipulado y/o de autor impostado creado deliberadamente para hacer daño. Este genérico denominado 'toxicidad' permite así englobar a las *fake news*, pero también a otros fenómenos como desinformación, discurso de odio, acoso, discriminación o *cyberbullying*, entre otros. Dicha toxicidad puede transmitirse mediante cualquier vía de comunicación, no obstante, se propaga preferiblemente por las redes sociales y mensajería instantánea, pues son los canales más efectivos para la difusión masiva (Casero-Ripollés *et al*., 2016). Por ello, plataformas como Twitter, concebidas como "la mejor herramienta de democratización que habían desarrollado nuestras sociedades", han pasado a ser consideradas, una década después de su nacimiento, "una gran amenaza para los sistemas democráticos occidentales" (Magallón-Rosa, 2019, p. 53).

Arraigada a esa transformación, la esfera pública digital también ha visto redefinido su significado dando entrada a la desconfianza y al cuestionamiento de la legitimidad. Bernard Williams explica en *Truth and Truthfulness* (2002) cómo el pensamiento contemporáneo se encuentra constantemente en alerta para no ser engañado, por lo que, la inclinación por conocer los entresijos de los hechos está en auge: "este interés en la veracidad da impulso a un proceso de crítica que debilita la confianza en que haya una verdad segura". Influye, igualmente, 'la paradoja del conspiranoico' (Elías, 2019), por la cual aquellas personas que prestan más atención activa a la manipulación proveniente de los medios de comunicación son más proclives a ser manipulados, pues rechazan las vías convencionales y tienden a interaccionar más a menudo con contenidos tóxicas, amplificando el alcance de la toxicidad en las conversaciones online.

Asimismo, el público es escéptico de la información proporcionada por los medios de comunicación debido a la polarización y los procesos engañosos como el *clickbait*, nacidos a raíz de la competencia comunicativa digital que surge en el entorno de las redes sociales. El resultado de este fenómeno Daniel Innerarity (2018) lo define como la 'uberización de la verdad', es decir, la "desprofesionalización del trabajo de la información" que ha debilitado los clásicos monopolios- universidad y prensa- en beneficio de las redes sociales. Al implantarse el modelo de usuarios prosumidores en una cultura participativa (Jenkins, 2006), se comienza a agrietar el monopolio de la agenda pública por parte de los medios y la credibilidad periodística dando lugar a la era de la posverdad.

En este contexto, lo importante no son los hechos, lo importante es la narración (Frankfurt, 2006), es decir, los hechos objetivos son menos influyentes en la formación de la opinión pública que las apelaciones a la emoción y las creencias (Keyes, 2004) que transmitimos en la propia narración. Al no dialogar con referencia a los hechos, los ciudadanos pierden la consciencia de cuál es la verdad. Al debatir sobre opiniones, se pervierte qué conforma la realidad verdadera y la realidad paralela creada por la posverdad. En definitiva, como expresa el filósofo norteamericano, Lee McIntyre, en su obra *Posverdad* (2018, p. 24): "si lo que se predica es la renuncia a valores como verdad (…) u objetividad, se despeja en gran parte el camino para imponer a los ciudadanos los intereses de quién miente", fomentando la propagación de mentiras, de odio y violencia a ciertos colectivos, es decir, facilitando la intoxicación de la comunicación *online*.

Asimismo, otro de los cambios fundamentales en el contexto social actual ha sido el auge de los sistemas algorítmicos como sistemas de curación de contenidos en las redes sociales. Las plataformas como Twitter se han convertido en los nuevos *gatekeepers*, empleando complejos algoritmos para seleccionar los contenidos en función de la comportamiento, participación y relevancia percibida de sus usuarios, alterando la forma en que fluye y se consume la información (Zuiderveen *et al*., 2018). Este control algorítmico determina qué contenido ven los usuarios con el objetivo efectivo de dar forma a sus experiencias online. Sin embargo, autores como Noble (2018) y Barocas *et al*. (2023) destacan los riesgos potenciales asociados con el *gatekeeping* algorítmico señalando que estos algoritmos pueden amplificar involuntariamente los contenidos tóxicos relacionados con rasgos raciales y de género.

1.2. El dilema por el control de la tecnología

En un estudio publicado en la revista científica Science en 2018, se analizaron 126.000 rumores difundidos por aproximadamente 3 millones de personas entre 2006 y 2017. Concluyeron que, "en términos de ritmos cotidianos de difusión, las noticias falsas tienen un 70% más de probabilidades de ser retuiteadas que las verdaderas". A una publicación verdadera "le lleva 6 veces más tiempo alcanzar 1.500 usuarios que a una falsa" y, si se trata de una cascada de RTs virales, "un hilo falso consigue una interacción entre 10 y 20 veces mayor que una cadena de hechos verdaderos" (Vosoughi *et al*., 2018, p. 1149). Además, los investigadores del MIT Lab señalaron el uso de tecnología y, concretamente, de herramientas de inteligencia artificial como una de las estrategias principales con la que se consiguieron viralizar los contenidos falsos. No obstante, la denominada "alta tecnología" (López-García y Vizoso, 2021) incide crucialmente tanto en la fase de digitalización de las redacciones como en los debates del periodismo actual.

El término, "inteligencia artificial" (IA), nació en 1956 durante la Conferencia de Darmouth, constituida *ad hoc* por veinte intelectuales para desarrollar el aprendizaje de máquinas. Precisamente, se le atribuye a John McCarthy, uno de los organizadores, la definición de IA. Según recoge una autorrecopilación de la Universidad de Stanford (2007, p.2), el propio autor denomina a la IA como "la ciencia y la ingeniería para fabricar máquinas inteligentes, especialmente programas informáticos inteligentes, (...) y comprender la inteligencia humana". Seis décadas después de su invención, Google ha concretado la conceptualización de McCarthy explicando cuáles son las características de la inteligencia humana que aspira a replicar. En 2018, durante la presentación de los principios y límites éticos de la IA para la compañía, su CEO, Sundar Pichai, resaltó la "resolución de problemas, creatividad y adaptabilidad" como las capacidades que "un mecanismo no humano" debe demostrar.

La descripción de McCarthy resalta la importancia del factor humano en la creación de máquinas, incorporando un elemento de "constructivismo social". Bijker y Pinch (1987) argumentan que las máquinas aprenden a partir de un corpus sesgado por las prácticas y valores culturales de la sociedad que las crea, lo que da el control a la acción humana. El avance tecnológico, por lo tanto, se ve influenciado por el cambio social, en línea con Kaplan (2009). Por otro lado, la visión del CEO de Google defiende el "determinismo tecnológico", donde las máquinas demuestran su aprendizaje de manera autónoma. McLuhan (1996) y Ellul (1962) también respaldan la idea de que la tecnología determina el cambio social. En tecnología, lo posible implica lo necesario; todo lo que esté alguna vez disponible, será necesariamente usado (Diéguez, 2005). Así es como la llegada de algoritmos personalizadores como Spotify ha sustituido en gran medida a las radios musicales o cómo los servicios de *streaming* han cambiado nuestras formas de consumo (*binge watching*) respecto a la televisión y al cine.

Igualmente, en los inicios de las redes sociales, los mensajes publicados no incidían en la agenda mediática; en cambio, hoy en día las plataformas han adquirido tanta relevancia que un post viral (Carral *et al*., 2023) puede abrir un telediario, cambiar resultados electorales o desencadenar manifestaciones en contra del poder político (Primaveras Árabes,15M en España, *gilets jaunes*). Por ello, muchos autores han cuestionado el papel de los *gatekeepers* humanos en un entorno de nuevos medios, donde el proceso de curación de contenidos en las redes sociales, motores de búsqueda o agregadores de noticias está dirigido por algoritmos cuyos criterios se forman en base a criterios individuales de los usuarios (Jürgens *et al*., 2011; Meraz y Papacharissi 2013). De esta manera, son los algoritmos quienes redistribuyen y canalizan la información promoviendo nuevos patrones en el flujo de las noticias (Wallace, 2017), dando forma a lo que se considera de interés público (Napoli, 2014) y participando en el proceso de construcción de la nueva realidad social (Just y Latzer 2016).

En el caso de las redes sociales, además, la mayoría de los espacios que ofrecen mecanismos de *gatekeeping* descentralizados están abiertos a todas las personas (Wallace, 2017). El control se ejerce una vez el comentario ya ha sido publicado y permanece a expensas de que la audiencia lo reporte en caso de ser ofensivo. Por ello, cualquier vertiente tóxica de la conversación nace parcialmente en consecuencia de ese proceso de post-moderación de la red social. No obstante, la acción humana acompaña también con sus conductas esta transición de la moderación humana a la tecnológica. En positivo, con la toma de decisiones como la creación de legislación o la invención de otros artefactos que reduzcan esos efectos perniciosos, y en negativo, fomentando y amplificando comportamientos incorrectos (acoso, insultos, falsedades, discriminación) en las redes sociales.

El discurso tóxico provoca que las personas sean menos propensas a participar en los espacios públicos digitales. Específicamente, este efecto silenciador impacta más a las voces marginadas de la sociedad como, por ejemplo, la comunidad LTGBIQ+ (Martínez, 2022). Los espacios de diálogo para la participación cívica -la sección de comentarios en artículos de medios de comunicación o los hilos de conversación en redes sociales y foros- se ven contaminados. De esta manera, toxicidad como las campañas de desinformación o los insultos fruto del elevado nivel de polarización ponen en peligro el periodismo independiente, conducen a bajadas en las ratios de fidelidad de los usuarios, desalientan a los anunciantes, incrementan los costes al medio por contratar moderadores o, llegado al extremo, provocan el cierre de esos espacios de debate impidiendo la participación ciudadana, el pluralismo ideológico y la comunicación bidireccional mediática (Fuchs, 2021).

En ese sentido, el área de distribución es la que más beneficiada se ha visto por la implementación de tecnología en el proceso comunicativo con mejoras en la personalización y recomendación de contenido, nuevos formatos de envío de contenido (*newsletters*) e interacción y asistencia a la audiencia (*chatbots*, *wizards*) (Sánchez-García *et al*., 2023). Sin embargo, la moderación de comentarios también está viviendo un período de innovación, pero de forma menos perceptible para la audiencia debido a su funcionamiento interno en la redacción. Herramientas de Inteligencia Artificial han sido creadas para eliminar el contenido tóxico de los comentarios y fomentar una conversación más sana. Por ejemplo, Perspective API colabora, en el ámbito del periodismo, con The New York Times, EL PAÍS o Le Monde, entre otros medios, creando un artefacto inteligente que encuentra patrones para detectar lenguaje abusivo y así calificar comentarios según la toxicidad. De esta manera, se agiliza la tarea de clasificación de comentarios que realiza el moderador profesional para que el medio de comunicación pueda actualizar la sección de comentarios en tiempo real a su audiencia prosumidora (Jigsaw, 2016).

Asimismo, la innovación ha llegado a otros mundos sociales como el *gaming*. FACEIT, la plataforma líder del sector incorporó también Perspective API para mejorar el trabajo de los moderadores humanos, al tiempo que fomentaba nuevas formas para que la comunidad interactúe entre sí libre de acoso, pero sin despojar a sus usuarios de su personalidad (Jigsaw, 2019b). Existe, no obstante, una vertiente de la moderación de contenidos visible para la audiencia y enfocada a las plataformas sociales como YouTube, Facebook, Twitter, Reddit y Disqus. Dado su sistema algorítmico descentralizador en la moderación, a través de otra herramienta de IA experimental denominada Tune, los usuarios cuentan con una extensión con la que pueden establecer el nivel de toxicidad de las conversaciones en las redes sociales (Jigsaw, 2019a).

2. Objetivos

En ese sentido, esta investigación trata de conocer el funcionamiento de dos herramientas de IA a través del estudio de un caso concreto como es la conversación política española en Twitter. Este caso es un mero ejemplo para testear los instrumentos, aún en fase experimental, no obstante, se podrían escoger temáticas tan variadas como el análisis de los medios de comunicación, de diferentes deportistas o clubs, otros argumentos discursivos en redes como los antivacunas o los negacionistas del cambio climático, e incluso, estudiar contextos temporalmente concretos como el 1-O en Cataluña o el 6 de enero de 2021 cuando se produjo el asalto al Capitolio en Estados Unidos.

Específicamente, se ha decidido elegir esta temática, primero, porque los líderes políticos junto con los periodistas son las figuras más influyentes en la definición de los límites y los contenidos que forman el debate público (Casero-Ripollés *et al*., 2016). En segundo lugar, porque Twitter no solo permite el estudio de los mensajes difundidos, sino también el análisis de los círculos de apoyo online creados para defender, promocionar o atacar (Chadwick, 2013). Y, en tercer lugar, por el interés en el estudio de la toxicidad a través de la herramienta Perspective. Un debate caracterizado por lenguaje ofensivo y argumentos falaces desalienta a la participación ciudadana y aumenta la polarización política de ideologías extremistas (Stryker *et al*., 2016).

Dado que el uso concreto de la herramienta Perspective en el mundo de la comunicación es un tema de tan reciente actualidad, apenas existen antecedentes exploratorios (Hosseini *et al*., 2017; Jain *et al*., 2018; Guerrero-Solé y Philippe, 200; Rieder y Skop, 2021) de los que se puedan extrapolar conclusiones acerca de la materia y que sirvan de sugestión a la hora de enunciar hipótesis. Por esta razón, se ha propuesto en su lugar dos preguntas exploratorias de investigación (PI) en relación con el objetivo anteriormente explicado:

- PI1 ¿Qué grado de toxicidad tiene la conversación sobre la política española en Twitter?
- PI2 ¿Cómo es la audiencia que conforma esas discusiones online?

3. Metodología

3.1. Objeto de estudio

Con el propósito de responder a estas incógnitas, se ha seleccionado para el estudio un período de tiempo perteneciente a la coyuntura de una conversación *online* ordinaria. Específicamente, se ha evitado la elección de un espacio temporal condicionado por eventos determinantes tales como una campaña electoral; ya que no sería adecuado extrapolar los resultados de un tiempo tan agitado e inaudito como una conclusión representativa de las dinámicas del debate diario. Asimismo, buscando la mayor neutralidad posible en la elección de los casos de estudio, se determinó analizar todos los perfiles en Twitter de los líderes de los partidos que hubieran obtenido más de un 1% de los votos en las elecciones generales en España del 10 noviembre 2019: Santiago Abascal (Vox), Inés Arrimadas (Cs), Ione Belarra (Unidas Podemos), Pablo Casado (PP), Íñigo Errejón (Más País-EQUO), Gabriel Rufián (ERC) y Pedro Sánchez (PSOE).

Por lo tanto, ante un escenario de temáticas y voces tan plurales, se optó por utilizar una metodología cuantitativa híbrida formada por procedimientos computacionales y manuales para el análisis de un objeto de estudio n=43.165 tuits. El corpus se compuso por los tuits publicados del 18 al 24 de octubre de 2021 correspondientes a siete políticos españoles y a la cascada de contestaciones de sus audiencias que componen la totalidad de cada hilo de conversación. Cabe aclarar que se decidió centrar el foco en un tuit diario, y en el que más respuestas obtiene porque el interés de la investigación no recae sobre el mensaje del político, sino en los comentarios de la audiencia en cada publicación. Estudios previos (Lampe y Johnston, 2005) demuestran que las contestaciones "aumentan las posibilidades de que se inicie una conversación y se cree una comunidad" de "usuarios comprometidos y motivados para futuras participaciones" (Burke *et al*., 2010, p. 4).

También se debe señalar que únicamente fueron analizados los mensajes publicados en castellano, ya que las herramientas de procesamiento de lenguaje natural utilizadas todavía no están disponibles para idiomas como el catalán o el vasco. Por este motivo técnico, Mertxe Aizpurua (EH Bildu), Laura Borrás (JuntxCat), Aitor Esteban (PNV) y Mireia Vehí (CUP) quedaron excluidos. Sin embargo, a pesar de que la recopilación y el primer análisis se hicieron con el soporte de diferentes herramientas, debido a su funcionamiento experimental en español, se ha necesitado la supervisión humana para corregir ciertos errores causados por los sesgos en Botometer y para suplir las limitaciones de la versión gratuita de Communalytic. Asimismo, se han llevado a cabo cálculos manuales en el cruce de variables analizadas por distintas herramientas.

3.2. Herramientas y técnicas para el análisis

Una vez seleccionada la muestra, se procedió a recopilar el material de estudio y se ejecutó el análisis de toxicidad a cada conjunto de datos, de nuevo, con la previa autorización para el uso de la API de Google, Perspective. Esta herramienta de IA se basa en modelos de aprendizaje automático con el fin de medir el nivel de toxicidad de una publicación y el impacto que podría tener en la conversación. La categorización y la capacidad de detección de la propia plataforma producen sesgos que los creadores de Perspective API han ido reduciendo desde su lanzamiento en 2016. Los falsos positivos y falsos negativos representan el problema prioritario, ya que "mensajes relacionados con la identidad reciben puntuaciones de toxicidad inapropiadamente altas o bajas". Términos como 'negro', 'musulmán', 'feminista', 'mujer o 'gay' "suelen tener puntuaciones más altas porque los comentarios sobre esos grupos están sobrerrepresentados en críticas abusivas y tóxicas" y, en sentido contrario, "hombre, "blanco" o "hetero" reciben puntuaciones más bajas (Jigsaw, 2021). Además, la herramienta cuenta con ciertas limitaciones: su carácter experimental en idiomas que no son inglés y la incapacidad para interpretar la ironía o la toxicidad específica como el argot.

En el siguiente paso se examinaron los usuarios que participaron en todos los hilos de conversación del objeto de estudio gracias a Botometer. El algoritmo de aprendizaje automático fue desarrollado en 2017 por el Observatorio de Redes Sociales (OSoMe por su nombre en inglés), perteneciente a la Universidad de Indiana. Está entrenado mediante la comparación de varios modelos de *machine learning*, cuyos algoritmos extraen las características (*features*) del perfil de la cuenta, los amigos, la estructura de la red social, los patrones de actividad temporal, el alcance, el idioma y el sentimiento del mensaje. Finalmente, Botometer otorga una puntuación entre 0 y 1 a cada usuario con el fin último de determinar si detrás de cada perfil se esconde una persona real (0) o un *bot* (1), aseverándolo con una probabilidad CAP (*complete automation probability*) del 80%. Es decir, de todas las cuentas señaladas como *bots* en la muestra, un 20% pueden llegar a ser falsos positivos. Dentro este porcentaje se incluyen casos como cuentas corporativas que publican a través de aplicaciones de escritorio, ya que el algoritmo puede llevar a confundir estas actuaciones con las típicas de un *bot*. Sin embargo, una verificación manual humana sobre la muestra ha conseguido reducir ese margen de error.

Sin embargo, una de las metas de este trabajo no se proyecta únicamente en la naturaleza de los usuarios, sino también en su función. Por este motivo, se ha decidido aplicar a estudio las seis categorías de análisis en español que ofrece Botometer, obteniendo así una puntuación entre 0 y 1 para cada una de ellas:

- Cámaras de eco (*eco-chamber bot*): cuentas que participan en grupos de seguimiento y comparten o eliminan contenido político en gran volumen.
- Seguidores falsos: *bots* comprados para aumentar el número de seguidores de un perfil.
- Financieros: *bots* que publican usando *cashtags*.
- Autodeclarado: *bots* creados a partir de la web botwiki.org.
- *Spammer*: cuentas que envían masivamente conjuntos de datos o información.
- Otros: como no se especifica bajo qué criterios se acepta señalar como *bots* a las cuentas, se ha decidido no trabajar con esta categoría para mantener la transparencia y la rigurosidad del análisis.

Tras confirmar la existencia de *bots* entre los usuarios participantes de las conversaciones desarrolladas a partir de los tuits que publican los líderes políticos, se ha procedido a estudiar cómo se relacionan estos robots automatizados con los usuarios no *bots*. Para ello, se han creado dos categorías para cada político: un ranking de los 10 usuarios con mayor probabilidad de ser *bots* de acuerdo con Botometer y otro top 10 de los usuarios que más han interactuado en sus hilos.

4. Resultados

A pesar de que autores como Moreno-López y Arroyo-López (2022) han estudiado previamente el tema de la toxicidad y los discursos del odio en redes, estos trabajos se han hecho mediante una metodología cualitativa como pueden ser entrevistas en profundidad y cuestionarios. Por ello, esta investigación aspira a revisar esos conceptos de manera cuantitativa y poniendo de relieve una nueva metodología con herramientas de IA. Para responder a la primera pregunta de investigación, se han analizado ciertas características negativas que afectan al contenido. Una segunda revisión más minuciosa nos ha permitido reconocer las particularidades de las categorías más relevantes a la hora de impedir una conversación online de buena calidad. Seguidamente, de cara a contestar la segunda pregunta de investigación, se ha procedido a la búsqueda de posibles *bots* que interfieran en la conversación real. Luego, esos datos se han confrontado con las estadísticas de los usuarios vertebradores de las múltiples discusiones nacidas a raíz del tuit del político. Con este procedimiento se aspira a averiguar posibles relaciones entre los *bots*, los prosumidores centrales y la toxicidad.

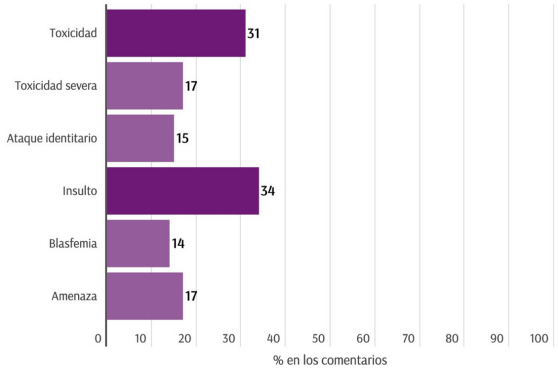
4.1. PI1: ¿Qué grado de toxicidad tiene la conversación política de España en Twitter?

Se define 'toxicidad' como la cualidad de un comentario de ser grosero, irrespetuoso o irrazonable que haga que otro usuario decida abandonar la conversación. Así describe el término la propia herramienta de IA utilizada en esta investigación, Perspective, creada para facilitar la conversación y evitar la toxicidad y el acoso online. No obstante, en español, Perspective ofrece cinco atributos más en los que categorizar los comentarios de los usuarios: toxicidad severa, cuando el comentario incita al odio o resulta agresivo; ataque identitario, dirigido expresamente por razones de identidad; insulto, entendido como comentario incendiario y negativo contra un individuo o un grupo; blasfemia, si se utiliza un lenguaje obsceno o profano; y amenaza, cuando se considera que hay intención de infligir dolor o violencia contra otro(s) usuario(s).

Por ello, en una primera fase de estudio nos hemos centrado en los atributos. Se debe resaltar que el análisis solo comprende unidades textuales, por lo que fotografías, gifs, emoticonos y contenido en audio y/o vídeo quedan fuera del alcance de inspección. El modelo de aprendizaje automático no detecta qué errores de deletreado o letras separadas por signos ortográficos puedan entenderse como un insulto, potenciando el sesgo de 'falsos negativos', como ya demostraron Jain *et al.*, (2018). Asimismo, se debe recordar que la herramienta ha sido entrenada con comentarios negativos en cualquier ámbito, no puramente políticos.

Si se analizan los seis atributos de forma aislada (Figura 1), los niveles de blasfemia (14%), ataques identitarios (15%), toxicidad severa (17%) y amenazas (17%) se encuentran por debajo de la media global. Sin embargo, si se focaliza la atención en el grado de toxicidad, el resultado se incrementa al 31% de comentarios tóxicos de media. En esa misma dirección ascendente, Perspective también parece señalar que los insultos son la categoría más preocupante, pues, de promedio, una de cada tres publicaciones contiene mensajes negativos e incendiarios contra un individuo o un grupo. Sin embargo, se debe tener en cuenta que los resultados podrían venir sesgados por la propia capacidad de detección del algoritmo. Es decir, puede tener menor sensibilidad para detectar una categoría que otra, llevándonos, en este caso, a sobreestimar la presencia de los insultos frente a otros tipos.

Figura 1: *Atributos que identifica Perspective y su presencia porcentual.*



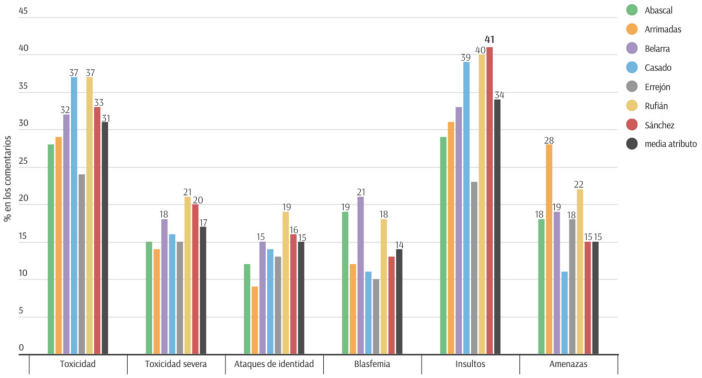
Fuente: Elaboración propia.

En una segunda fase de análisis se han comparado las medias de los atributos de cada político individualmente. Así se vislumbra una tendencia porcentual al alza, pero coherente con las dinámicas predispuestas en la anterior etapa. De hecho, hasta en seis de los siete perfiles analizados, el atributo con

el promedio más alto se vincula con la aparición de insultos en las respuestas de los usuarios al tuit inicial, como sugería Perspective. Y en el caso excepcional, el de Errejón, su promedio máximo se asocia a la toxicidad, el segundo atributo con un promedio mayor. Como observamos en la Figura 2, las métricas a nivel individual indican una propensión similar a las tendencias conjuntas, aunque cabe resaltar dos particularidades llamativas.

En primer lugar, las conversaciones que se producen a raíz de los tuits del líder de ERC, Gabriel Rufián, superan las medias globales en cualquiera de los seis atributos. Incluso, en la categoría de insultos, si la media conjunta alcanzaba el 34%, la suya personal se incrementa hasta el 40%. En segundo lugar, las estadísticas del presidente del Gobierno también sobrepasan a las medias conjuntas en cuatro de los seis atributos. Concretamente, el perfil de Pedro Sánchez recoge el promedio porcentual más alto (41%) de cualquier atributo, con algunos días en los que el 50% de los mensajes, es decir la mitad de todas las respuestas a su tuit, contienen insultos.

Figura 2: *Presencial porcentual de cada atributo en función de cada.*



Fuente: Elaboración propia.

En relación con este atributo, al explorar los puntos críticos, es decir, los índices de máximo nivel de cualquiera de los seis atributos, se consolida la enunciación de Perspective sobre los improperios. El porcentaje más alto de comentarios negativos en hasta seis de los siete sujetos analizados está vinculado a la aparición de insultos. Además, la proporción de agravios máxima sufrida por cada político es en todos los casos superiores a la media de la categoría, establecida en el 34%.

De nuevo, destacan Rufián, Sánchez y Casado, cuyos porcentajes son alarmantes. En hasta dos días de la semana estudiada, el 50% de los comentarios a propósito del tuit inicial tanto del líder de la oposición en ese momento como del presidente del Gobierno fueron considerados insultos por Perspective. La excepción recae en Belarra, cuyo pico máximo se asocia a la categoría de 'amenazas'. La líder de Podemos recibió comentarios intimidatorios en el 40% de las respuestas a su tuit, en el que ponía en valor las disculpas pronunciadas por Arnaldo Otegui en referencia a los crímenes de ETA. Sin embargo, la temática de la publicación de Belarra encaja perfectamente con los mensajes del resto de líderes políticos con más comentarios negativos.

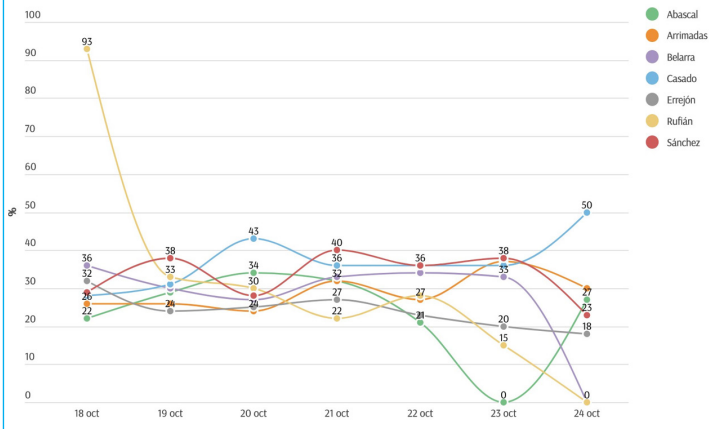
Una vez que nuestros resultados parecen confirmar el serio problema que suponen los insultos en la conversación sobre la política española en Twitter, tal y como señalaba Perspective; se han examinado pormenorizadamente los datos referentes al atributo de 'toxicidad'. Siguiendo la tendencia que marcaban los indicadores globales, podemos agrupar a algunos líderes políticos en función del comportamiento de la audiencia. Por un lado, el análisis del grado de toxicidad señala que Arrimadas (29%), Abascal (28%) y Errejón (24%) se sitúan por debajo del promedio de comentarios tóxicos (31%). Incluso, el líder de Vox y el de Más País-EQUO cuentan en sus publicaciones más negativas con los porcentajes más bajos de entre todos los líderes, con un 34% y 32% de toxicidad.

Por otro lado, Rufián encabeza el grupo de los políticos con los hilos de conversación más tóxicos. Sus tuits cosechan de media un 37% de toxicidad, aunque cabe resaltar que, en su pico máximo, Perspective consideró que hasta 93 de cada 100 comentarios fueron tóxicos. Es relevante tener en cuenta la cuantía de publicaciones sobre las que se han hecho estas estimaciones. De esa manera, se puede percibir cómo los perfiles de Rufián y de Belarra son los que más respuestas reciben, superando las 10.000 contestaciones. Casado y Sánchez obtienen la mitad; Abascal y Errejón, tres veces menos y Arrimadas, solo una décima parte. Teniendo en cuenta estos datos, se observa una relación paralela en las conductas de la audiencia hacia

Rufián y Belarra. Ambos tienen el promedio personal sobre el grado de toxicidad por encima de la media (Figura 3) y, además, los dos cuentan con los mínimos de toxicidad más altos: en sus publicaciones menos tóxicas, el volumen de comentarios negativos no desciende del 27%.

Por su parte, Casado obtiene, como Rufián, un grado de toxicidad promedio del 37% y su tuit más convulso también supera la media grupal de toxicidad. De hecho, una de cada dos respuestas en la publicación más álgida del popular contenía toxicidad. Sin embargo, los resultados de Casado manifiestan una distribución estadística normal, a diferencia de los de Rufián que resultan en una asimetría positiva. Mientras que para el político catalán hay una diferencia de 78 puntos porcentuales entre su toxicidad mínima y la máxima, para el popular el rango se reduce a los 22 puntos. Este dato muestra cómo Casado cuenta con un grado de toxicidad elevado, pero estable y continuado en el tiempo, mientras que Rufián sufre sacudidas tremendamente tóxicas (máximo de toxicidad: 93%) alternadas con otras rachas bastante más calmadas (mínimo: 15%).

Figura 3: *Evolución del grado de toxicidad en cada tuit de cada líder político.*



Fuente: Elaboración propia.

4.2. PI2: ¿Cómo es la audiencia que conforma las discusiones online?

En la primera pregunta de investigación se ha estudiado cómo se comporta la audiencia frente a los mensajes de los políticos. A continuación, se procederá a conocer quién es esa audiencia prosumidora, es decir, quién es ese público consume los mensajes al igual que participa en la conversación produciendo otros tuits. Con ese objetivo, se ha cuestionado si son usuarios reales o, por el contrario, se trata de bots programados para difundir contenido hiriente o confuso; cómo se relacionan entre sí y en qué punto de la conversación se crea más toxicidad.

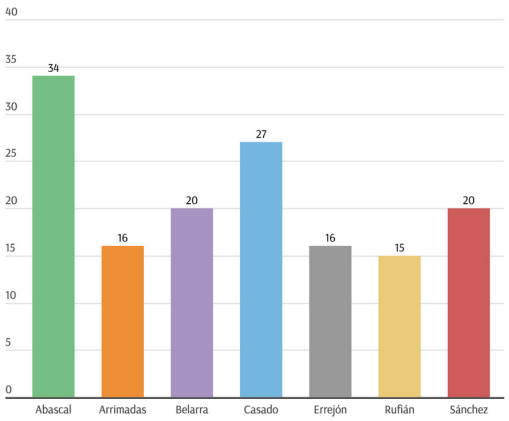
Con una probabilidad CAP (*complete automation probability*) otorgada por el propio algoritmo Botometer del 80%, las conversaciones estudiadas entre el lunes 18 de octubre y el domingo 24 de octubre de 2021 contaban de media con un 21% de usuarios *bots*. No existe una definición universalmente acordada para designar qué es un bot debido a la amplia gama de comportamientos que pueden tener. No obstante, los creadores de Botometer (Yang *et al*., 2020) determinan que un *bot* es una "cuenta de redes sociales controlada, al menos en parte, a través de *software*". Pese a que hay muchos tipos de *bots*, los maliciosos se pueden utilizar para manipular a los usuarios de redes sociales amplificando la desinformación, creando la apariencia de que algunas personas o ideas son más populares, cometiendo fraude financiero o difundiendo spam, entre otros (Schuchard *et al*., 2019). Por ello, se ha decidido estudiar tanto la existencia de *bots* como también la función que puedan ejercer gracias a las seis categorías de análisis en español que ofrece Botometer.

De esta manera, se ha constatado que Abascal es el político con más presencia de posibles bots en sus conversaciones (promedio del 34%). En segundo lugar, se sitúa Casado, con una media de 27% de bots participando en sus hilos (Figura 4). Y, empatados con un promedio del 20%, Sánchez y Belarra. La líder de Podemos, sin embargo, encabeza el ranking entre los *bots* autodeclarados, ya que una de cada cinco cuentas que le comentan son *bots* creados a través del

> "El 50% de los comentarios a propósito del tuit inicial tanto del líder de la oposición en ese momento como del presidente del Gobierno fueron considerados insultos por Perspective"

proyecto botwiki.org. Todos los demás líderes políticos cuentan con una media de *bots* autodeclarados más baja, pero ninguna inferior al 12% que ostenta Errejón. De hecho, al igual que sucedía con los niveles de toxicidad, el líder de Más País-EQUO apenas despierta relevancia entre la audiencia de *bots* (16%). Al contrario de lo que sucede con Rufián quien, a pesar de sus altos grados de toxicidad, posee la media más baja con un 15% de *bots* entre su público.

Figura 4: *Número de bots que comentan los tuits de cada político.*



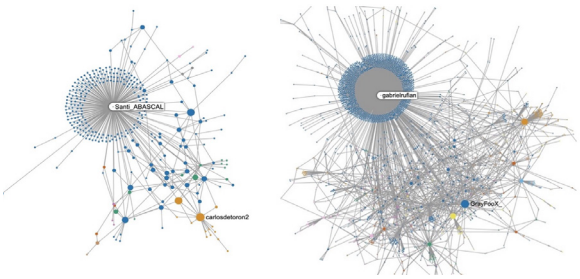
Fuente: Elaboración propia.

En ese sentido, resulta paradójico ver cómo cambian las tornas. Por una parte, Rufián también es el líder con menos seguidores falsos (3%), el único por debajo de la media del 5%. Enfrente, Arrimadas y Sánchez, con un 7% y 6% respectivamente; lo que indica que la líder de Ciudadanos cuenta con casi 50.000 seguidores *bots* y al presidente del Gobierno le siguen 96.000 usuarios *fakes*. Por otra parte, el líder de Vox que apenas destacaba en los análisis de toxicidad, vuelve a ser el primero, esta vez, alcanzando el pico máximo de *bots*. El 22 de octubre, hasta un 86% de los usuarios que comentaban en el tuit de Abascal, según Botometer, podrían considerarse *bots*. También Casado en su tuit del 18 de octubre obtuvo un 81% de bots entre los usuarios que le comentaron. Son reseñables estos dos ejemplos ya que, en comparación, el máximo de bots de los demás políticos nunca superó el 50%.

Por último, pese a que Abascal posee una cantidad mayor de *bots* entre quienes le comentan, Casado es el líder con más *bots* que funcionan como cámaras de eco. Un 22% de los perfiles que le responden o participan en las conversaciones nacidas a raíz de su hilo tienen la intención de compartir o eliminar contenido político a gran escala. Este dato supone que el líder popular cuenta con más del doble de *bots eco-chamber* que Sánchez (11%) o, por ejemplo, siete veces más que Arrimadas (3%). En términos generales, se ha concluido que los 10 usuarios más participativos de cada político no se identifican con un perfil *bot*, dicho de otra manera, que no existe ninguna evidencia por la cual se pueda afirmar que las cuentas *bots* son las que más comentan en las conversaciones creadas a raíz de los tuits publicados por políticos.

Asimismo, dado que en perfiles como el de Casado los *bots eco-chamber* suponen hasta un 22%, nos hemos preguntado si su función de compartir o eliminar contenido político en gran volumen puede involucrar contenido tóxico. Combinando los datos obtenidos de las dos herramientas de IA, Perspective y Botometer, se ha investigado si son *bots* quienes publican los mensajes más tóxicos. Cabe resaltar que se han elegido, por una parte, los 10 usuarios que Botometer considera con mayor probabilidad *bots* y, por la otra, las contestaciones al tuit que contengan al menos un 80% de toxicidad. Como resultado, tan solo el 9% de las estas publicaciones tóxicas fueron emitidas por alguna cuenta *bot* incluida en los top 10 de cada político.

Figura 5: *Interacciones en el estado 0 de Abascal y de Rufián.*



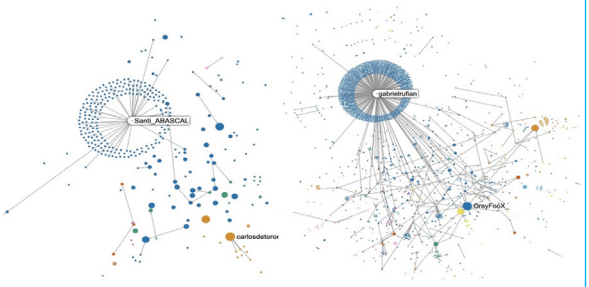
Fuente: Elaboración propia.

Por último, como en la primera pregunta de investigación se estudiaba a grandes rasgos si la audiencia emite comentarios tóxicos hacia el tuit publicado por el político, en esta segunda pregunta se ha tratado de ahondar en la cuestión averiguando dónde se crea más toxicidad. Para responder a esta pregunta se han tomado dos momentos como referencia y se han comparado. El primero equivale al estado natural (estado 0) de la conversación, midiendo todas las reacciones hacia el político y las interacciones entre los usuarios. El segundo viene dado por el grado de toxicidad que queremos estudiar. Dado que hemos

trabajado con un nivel tóxico del 80% en el resto de las cuestiones, hemos escogido también ese porcentaje para la comparación.

De esta manera, como se observa en la Figura 5 de los grafos de Abascal y de Rufián, en un primer momento (estado 0) se producen relaciones de toxicidad- por ínfimo que sea el grado- tanto en los comentarios dirigidos directamente al líder político como en las conversaciones que nacen de la interacción entre usuarios. De hecho, se pueden ver cómo se han creado ciertos clusters, una concentración de usuarios interconectados en la conversación de segundo y siguientes niveles, no en el hilo principal. En el grafo de Abascal destaca, por ejemplo, @carlosdetoron2, que capitanea una pequeña comunidad además de expresarse individualmente y de forma directa hacia el líder de Vox. En el grafo de Rufián, @GrayFooX_ lleva la voz cantante de las conversaciones secundarias, aunque también se aprecian a otros usuarios generando pequeños clústeres (en color amarillo, en naranja, en verde, en rosa).

Figura 6: *Interacciones con al menos un 80% de toxicidad de Abascal y Rufián.*



Fuente: Elaboración propia.

En cambio, cuando se acota la participación de los usuarios a relaciones con al menos un 80% de toxicidad en los comentarios, se distingue claramente que el contenido más tóxico se focaliza en mensajes directos al líder que publicó el tuit. La mayoría de las interacciones entre usuarios desaparecen, puesto que su retroalimentación no contiene niveles de toxicidad tan altos. No obstante, se mantienen los adalides de cada clúster, aun cuando no sean estos usuarios quienes emitan comentarios tóxicos, ya que al ser los centralizadores de las conversaciones secundarias, se les incluye en la discusión —al igual que al político— por parte de quienes publican de forma tóxica. Así vemos en la Figura 6 como @carlosdetoron2 conserva ciertas interacciones con un 80% o más de toxicidad en el hilo de Abascal e igual sucede con @GrayFoox a raíz del tuit de Rufián.

5. Discusión y Conclusiones

Visto lo anterior, esta investigación muestra cómo nuevas metodologías de IA pueden, primero, contabilizar un término hasta ahora tan abstracto como la toxicidad y, segundo, ayudan a poner claridad en un tema tan profundo y desconocido como es la naturaleza y los comportamientos de la audiencia en las redes sociales.

Para responder a la PI-1 sobre el grado de toxicidad que tiene la conversación sobre la política española en Twitter ha sido necesario realizar un estudio en profundidad, donde los resultados proporcionados por Perspective revelan que la conversación sobre temas políticos con Twitter está circunscrita en un entorno tóxico, donde, por ejemplo, alrededor de una de cada tres respuestas son insultos (34%). Además, se puede confirmar que se trata de un rasgo intrínseco a la conversación. En los datos de cada líder político, las respuestas que contienen insultos de los usuarios al tuit inicial son la categoría más popular en hasta en seis de los siete perfiles analizados, es decir, no estamos ante un señalamiento específico a ciertos políticos por su género o pertenencia a una ideología concreta. Sin embargo, cabe puntualizar la existencia de dos tendencias en la prevalencia de este fenómeno: existen perfiles como Casado, cuya conversación cuenta con un grado de toxicidad elevado, estable y continuado en el tiempo, mientras las de Rufián tan solo sufren sacudidas con un alto grado de toxicidad en momento puntuales. En ese sentido una posible línea de investigación futura podría ser el estudio de los motivos por los cuales se producen estas conductas, analizando si los patrones son debidos al sujeto o si la temática suscita el mismo impacto en otros actores.

Igualmente, este trabajo de investigación cuestionaba en su PI-2 cómo es la audiencia que conforma esas discusiones online. Los resultados indican la existencia de usuarios ficticios en la conversación, ya que, todos los hilos de Twitter analizados contaban, de promedio, con un 21% de usuarios *bots*. Sin embargo, se muestra que no existe ninguna asociación entre los *bots* más activos en las conversaciones de cada político y las publicaciones más tóxicas, ya que solo el 9% de las respuestas tóxicas fueron emitidas por bots. Por consiguiente, estos

hallazgos apoyan la misma teoría que estudios previos como Howard *et al*. (2017), quienes sostienen que son los bots quienes difunden originalmente más contenido tóxico, pero a la vez es la gente quien comparte en mayor medida las publicaciones. Por el contrario, los resultados de esta investigación contradicen parcialmente los hallazgos previos (Caldarelli *et al*., 2020; Shao *et al*., 2018). No se niega la capacidad difusora de estas cuentas, pero sí se demuestra que no son los bots, sino los usuarios humanos quienes más se involucran con los contenidos con un mayor nivel de toxicidad, ampliando así su alcance.

Asimismo, la granularidad de los datos que aportan estas herramientas de IA permite indagar individualmente cada sujeto de la muestra hasta poder conocer, por ejemplo, que Abascal es el político con más presencia de *bots* (promedio del 34%) y que también el pico máximo de *bots* se registró en una de sus conversaciones, con un 86% de usuarios falsos. De esta manera, el estudio de los *bots* puede resultar beneficioso y útil para ámbitos como la comunicación política, especialmente en la creación de estrategias de *marketing*. Además, el núcleo de investigadores académicos interesados en las burbujas de filtro puede servirse de estas herramientas para analizar la naturaleza de los usuarios que componen ese reducto. En este caso de estudio se observa, por ejemplo, como Casado es el líder con más *bots* (22%) que funcionan como cámaras de eco.

Por último, en todas las conversaciones se ha observado una misma tónica general: los usuarios comentaristas dirigen la toxicidad hacia el líder que publica el tuit a través de comentarios directos. Cuanto más se eleva el grado de toxicidad, menos interacciones se mantienen entre usuarios. Este patrón apunta a, por un lado, una descarga de comentarios negativos directamente hacia el político y, por otro, a una conversación de mejor calidad entre pares. Con respecto al primer comportamiento, la descarga de comentarios negativos directamente hacia el político, se deben tener en consideración las costumbres y contextos culturales de cada civilización. Al contrario del mundo occidental, estudios como el de Feldman (2023) muestran que, en Oriente, concretamente en Japón, los comentarios degradantes de los líderes japoneses rara vez se dicen cara a cara al destinatario, sino frente a audiencias generales que no son el objetivo de la degradación.

En definitiva, estas herramientas de IA, aún en fase experimental, ayudan a cuantificar nuevas realidades en las redes sociales. Este caso es un mero ejemplo para testear los instrumentos, no obstante, se podrían escoger temáticas tan variadas como el análisis de los medios de comunicación, de diferentes deportistas o clubs, otros argumentos discursivos en redes como los antivacunas o los negacionistas del cambio climático, e incluso, estudiar contextos temporalmente concretos como el 1-O en Cataluña o el 6 de enero de 2021 cuando se produjo el asalto al Capitolio en Estados Unidos. Sin embargo, el pilar fundamental de esta investigación reside en dar a conocer una nueva metodología para el estudio académico en y de las redes sociales a través de nuevas variables como la toxicidad – en todas sus vertientes, y ahondar en la divulgación sobre los algoritmos bots y su conexión con dicha toxicidad. Asimismo, se ha puesto de manifiesto la versatilidad tanto para la moderación de contenido en medios de comunicación como para las redes sociales.

6. Referencias

Barocas, S., Hardt, M. y Narayanan, A. (2023). *Fairness and machine learning: Limitations and Opportunities*. MIT Press.

Bijker, W. E. y Pinch, T. (1987). *The social construction of facts and artifacts*. https://acortar.link/Y4RqvA

Blanco-Alfonso I., García-Galera C. y Tejedor-Calvo S. (2019). El impacto de las fake news en la investigación en Ciencias Sociales. Revisión bibliográfica sistematizada. *Historia y Comunicación Social, 24*(2), 449-469. https://doi.org/10.5209/hics.66290

Burke, M., Kraut, R. y Joyce, E. (2010). Membership claims and requests: Conversation-level newcomer socialization strategies in online groups. *Small group research, 41*(1), 4-40. https://doi.org/10.1177/1046496409351936

Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M. y Saracco, F. (2020). The role of bot squads in the political propaganda on Twitter. *Commun Phys*, 3. https://doi.org/10.1038/s42005-020-0340-4

Carral, U., Tuñón, J. y Elías, C. (2023). Populism, cyberdemocracy and disinformation: analysis of the social media strategies of the French extreme right in the 2014 and 2019 European elections. *Humanit Soc Sci Commun 10*, 23. https://doi.org/10.1057/s41599-023-01507-2

Casero-Ripollés, A., Feenstra, R. A. y Tormey, S. (2016). Old and New Media Logics in an Electoral Campaign: The Case of Podemos and the Two-Way Street Mediatization of Politics. *The International Journal of Press/Politics, 21*(3), 378-397. https://doi.org/10.1177/1940161216645340

Chadwick, A. (2013). *The hybrid media system: Politics and power*. Oxford University Press

Diéguez, A. (2005). El determinismo tecnológico: indicaciones para su interpretación. *Argumentos de Razón Técnica*, 8, 67-87. http://www-formal.stanford.edu/jmc/whatisai.pdf

Elías, C. (2019). *Science on the Ropes. Decline of Scientific Culture in the Era of Fake News*. Springer-Nature. https://doi.org/10.1007/978-3-030-12978-1

Ellul, J. (1962). The Technological Order. *Technology and Culture, 3*(4), 394-421. https://doi.org/10.2307/3100993

Feldman, O. (2023). Challenging Etiquette: Insults, Sarcasm, and Irony in Japanese Politicians' Discourse. En O. Feldman (Ed.), *Political Debasement. The Language of Politics*. Springer. https://doi.org/10.1007/978-981-99-0467-9_5

Frankfurt, H. (2006). *On Bullshit: sobre la manipulación de la verdad*. Paidós

Fuchs, C. (2021). *Social media: A critical introduction*. Sage.

Goldhaber, M. H. (1997). The attention economy and the Net. *First Monday, 2*(4). https://doi.org/10.5210/fm.v2i4.519

Guerrero-Solé, F. y Philippe, O. (2020). La toxicidad de la política española en Twitter durante la pandemia de la COVID-19. *Hipertext.net*, 21, 133-139. https://doi.org/10.31009/hipertext.net.2020.i21.12

Hosseini, H., Kannan, S., Zhang, B. y Poovendran, R. (2017). *Deceiving Google's perspective API built for detecting toxic comments*. Cornell University. htttps://doi.org/10.48550/arXiv.1702.08138

Howard, P. N., Bradshaw, S., Kollanyi, B. y Bolsolver, G. (2017). Junk News and Bots during the French Presidential Election: What Are French Voters Sharing Over Twitter in Round Two? *ComProp data memo, 21*(3). https://acortar.link/IUZfrN

Innerarity, D. (2018, 31 diciembre). El año de la volatilidad. *El País*. https://elpais.com/elpais/2018/12/28/opinion/1546021545_365361.html

Jain, E., Brown, S., Chen, J., Neaton, E., Baidas, M., Dong, Z. y Artan, N. S. (2018, diciembre). *Adversarial Text Generation for Google's Perspective API*. 2018 International Conference on Computational Science and Computational Intelligence (CSCI), (pp. 1136-1141). IEEE. http://doi.org/10.1109/CSCI46756.2018.00220

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. NYU Press.

Jigsaw. (2016, septiembre 19). *New York times and Jigsaw partner to scale moderation platform*. Medium. https://acortar.link/t1R0hG

Jigsaw. (2019a, marzo12). *Tune: Control the comments you see*. Medium. https://acortar.link/ACX5fU

Jigsaw. (2019b, octubre 23). *One of Europe's largest gaming platforms is tackling toxicity with machine learning*. Jigsaw. https://acortar.link/KgcA2i

Jigsaw. (2021, febrero 10). *Helping authors understand toxicity, one comment at a time*. Medium. https://bit.ly/3KgjnSz

Jürgens, P., Jungherr, A. y Schoen, H. (2011). Small worlds with a difference: New gatekeepers and the filtering of political information on Twitter. *Proceedings of the 3rd international web science conference*. https://doi.org/10.1145/2527031.2527034

Just, N. y Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, culture & society, 39*(2), 238-258. https://doi.org/10.1177/0163443716643157

Kaplan, D. M. (Ed.). (2009). R*eadings in the Philosophy of Technology*. Rowman & Littlefield Publishers.

Keyes, R. (2004). *The Post-Truth Era: Dishonesty and Deception in Contemporary Life*. St. Martin's Press.

Lampe, C. y Johnston, E. (2005). Follow the effects of feedback on new members in an online community. *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*. https://doi.org/10.1145/1099203.1099206

López-García, X. y Vizoso, Á. (2021). Periodismo de alta tecnología: signo de los tiempos digitales del tercer milenio. *Profesional de la Información, 30*(3). https://doi.org/10.3145/epi.2021.may.01

Magallón-Rosa, R. (2019). *Unfaking news: cómo combatir la desinformación*. Pirámide.

Martínez Valerio, L. (2022). Mensajes de odio hacia la comunidad LGTBIQ+: análisis de los perfiles de Instagram de la prensa española durante la "Semana del Orgullo". *Revista Latina de Comunicación Social*, 80, 364-388. https://doi.org/10.4185/RLCS-2022-1749

McCarthy, J. (2007). *What is Artificial Intelligence?* Stanford University.

McIntyre, L. (2018). *Post-truth*. MIT Press.

McLuhan, M. (1996). *El medio es el masaje. Un inventario de efectos*. Paidós.

Meraz, S. y Papacharissi, Z. (2013). Networked gatekeeping and networked framing on# Egypt. *The international journal of press/politics, 18*(2), 138-166. https://doi.org/10.1177/1940161212474472

Moreno-López, R. y Arroyo-López, C. (2022). Redes, equipos de monitoreo y aplicaciones móvil para combatir los discursos y delitos de odio en Europa. *Revista Latina de Comunicación Social*, 80, 347-363. https://doi.org/10.4185/RLCS-2022-1750

Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication theory, 24*(3), 340-360. https://doi.org/10.1111/comt.12039

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Pichai, S. (2018, 7 de junio). AI at Google: Our Principles. *Blog Google.* https://blog.google/technology/ai/ai-principles/

Rieder, B. y Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society, 8*(2), https://doi.org/10.1177/2053951721104618

Sánchez-García, P., Merayo-Álvarez, N., Calvo-Barbero, C. y Diez-Gracia, A. (2023). Spanish technological development of artificial intelligence applied to journalism: companies and tools for documentation, production, and distribution of information. *El Profesional de la Información, 32*(2). https://doi.org/10.3145/epi.2023.mar.08

Schuchard, R., Crooks, A., Stefanidis, A. y Croitoru, A. (2019). Bots fired: examining social bot evidence in online mass shooting conversations. *Palgrave Communications, 5*(1), 1-12. https://doi.org/10.1057/s41599-019-0359-x

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K-C, Flammini A. y Menczer, F. (2018). The spread of low-credibility content by social bots. *Nat Commun*, 9. https://doi.org/10.1038/s41467-018-06930-7

Silverman, C. (2014). *Verification Handbook*. European Journalism Centre.

Southwell, B. G., Thorson, E. A. y Sheble, L. (Eds.). (2018). *Misinformation and mass audiences*. University of Texas Press.

Stryker, R., Conway, B. A. y Danielson, J. T. (2016). What is political incivility? *Communication Monographs, 83*(4), 535-556. https://doi.org/10.1080/03637751.2016.1201207

Toffler, A. (1975). *Alvin Toffler*. Pacifica Tape Library.

Tune. (s/f). *Tune experimental*. Google.com. https://acortar.link/bvJqZk

Vosoughi, S., Roy, D. y Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wallace, J. (2017). Modelling contemporary gatekeeping: The rise of individuals, algorithms, and platforms in digital news dissemination. *Digital Journalism, 6*(3), 274-293. https://doi.org/10.1080/21670811.2017.1343648

Wardle, C. y Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Strasbourg: Council of Europe. https://acortar.link/dsM2G5

Yang, K. C., Varol, O., Hui, P. M. y Menczer, F. (2020). Scalable and generalizable social *bot* detection through data selection. *Proceedings of the AAAI conference on artificial intelligence, 34*(1), 1096-1103. https://doi.org/10.1609/aaai.v34i01.5460

Zuiderveen, F. B., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H. y Helberger, N. (2016). Should we worry about filter bubbles? Internet Policy Review. *Journal on Internet Regulation, 5*(1). https://doi.org/10.14763/2016.1.401

CONTRIBUCIONES DE AUTORES/AS, FINANCIACIÓN Y AGRADECIMIENTOS

Contribuciones de los/as autores/as:

Conceptualización: Carral, Uxía y Elías, Carlos. **Análisis formal:** Carral, Uxía y Elías, Carlos. **Curación de datos:** Carral, Uxía. **Redacción-Preparación del borrador original:** Carral, Uxía y Elías, Carlos. **Redacción-Revisión y Edición:** Elías, Carlos. **Visualización:** Carral, Uxía y Elías, Carlos. **Supervisión:** Elías, Carlos. **Administración de proyectos:** Elías, Carlos. **Todos los/as autores/as han leído y aceptado la versión publicada del manuscrito:** Carral, Uxía y Elías, Carlos.

Financiación: El presente texto nace en el marco del dos proyectos concedidos a miembros de la Universidad Carlos III de Madrid de los cuales recibe financiamiento esta investigación: 'Pseudociencia, teorías conspirativas, *fake news* y alfabetismo mediático en la comunicación en salud' (Ref: PID2022-142755OB-I00) dentro del Plan Nacional 'Proyectos de Generación de Conocimiento' del Ministerio de Ciencia e Innovación y 'UE, desinformación y *fake news*' (Ref: 610538-EPP-1-2019-1-ES-EPPJMO-CHAIR), una Cátedra Jean Monnet de la Comisión Europea.

Agradecimientos: Los autores desean agradecer sinceramente el valioso asesoramiento a Rubén Míguez Pérez, Doctor en ingeniería de Telecomunicaciones por la Universidad de Vigo, que hizo también posible la realización de esta investigación.

AUTOR/A/ES:

Uxía Carral Vilar
Universidad Carlos III de Madrid. España.

Estudiante predoctoral en Investigación en Medios de Comunicación en la Universidad Carlos III de Madrid (UC3M). Sus líneas de investigación son las redes sociales, la innovación en la comunicación y el humanismo tecnológico. Ha hecho estancias en University Technology of Sydney (UTS) y en La Sapienza (Roma, Italia). También ha trabajado como periodista en medios de comunicación como Cadena SER y de verificación como Newtral. Además, ha llevado la comunicación de varios proyectos europeos financiados por la Comisión Europea y ha colaborado como periodista de visualización de datos en el Círculo de Análisis Euromediterráneo (CAEM). Autora de varios libros y capítulos sobre desinformación, populismos, redes sociales y Unión Europea.
ucarral@hum.uc3m.es

Índice H: 4
Orcid ID: https://orcid.org/0000-0002-2329-3331
Google Scholar: https://scholar.google.es/citations?user=ART9ru0AAAAJ&hl=es&oi=ao
ResearchGate: https://www.researchgate.net/profile/Uxia-Carral-2
Academia.edu: https://uc3m.academia.edu/UxiaCarral

Carlos Elías Pérez
Universidad Carlos III de Madrid. España.

Síntesis del currículo del autor/a de 120 palabras.

Catedrático de Periodismo de la Universidad Carlos III de Madrid (UC3M) y catedrático Jean Monnet "UE, desinformación y fake news". Se especializó en ciencia, tecnología y esfera pública en estancias postdoctorales de un año en London School of Economics y otro año en Harvard. Ha trabajado como periodista en la Agencia

Efe (política) y El Mundo (responsable de ciencia). Su último libro es Science on the Ropes. Decline of Scientific Culture in the era of Fake News (Springer, 2019). Es director del máster de Comunicación Corporative e Institucional de la UC3M y colabora habitualmente con medios como El Mundo, RNE y RTVE.
carlos.elias@uc3m.es

Índice H: 20
Orcid ID: https://orcid.org/0000-0002-1330-4324
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=31267503700
https://www.scopus.com/authid/detail.uri?authorId=57693924900
Google Scholar: https://scholar.google.es/citations?user=jiCVGcIAAAAJ&hl=es
ResearchGate: https://www.researchgate.net/profile/Carlos-Elias-3

Revista Latina de Comunicación Social

ARTÍCULOS RELACIONADOS

Aramburú Moncada, L. G., López Redondo, I., & López Hidalgo, A. (2023). Inteligencia artificial en RTVE al servicio de la España vacía. Proyecto de cobertura informativa con redacción automatizada para las elecciones municipales de 2023. *Revista Latina de Comunicación Social*, 81, 1-16. https://doi.org/10.4185/RLCS-2023-1550

Deliyore Vega, M. D. (2021). Redes como espacio de comunicación para la educación virtual de estudiantes con discapacidad en Costa Rica en tiempos de pandemia. *Historia y Comunicación Social, 26*(Especial), 75-85. https://doi.org/10.5209/hics.74243

Demuner Flores, M. del R. (2021). Uso de redes sociales en microempresas ante efectos COVID-19. *Revista de Comunicación de la SEECI*, 54, 97-118. https://doi.org/10.15198/seeci.2021.54.e660

Hueso Romero, J. J. (2022). Creación de una red neuronal artificial para predecir el comportamiento de las plataformas MOOC sobre la agenda 2030 y los objetivos para el desarrollo sostenible. *Vivat Academia. Revista de Comunicación*, 155, 61-89. https://doi.org/10.15178/va.2022.155.e1386

Sancho Escrivá, J. V., Fanjul Peyró, C., De la Iglesia Vayá, M., Montell, Joaquín A., & Escartí Fabra, M. J. (2020). Aplicación de la inteligencia artificial con procesamiento del lenguaje natural para textos de investigación cualitativa en la relación médico-paciente con enfermedad mental mediante el uso de tecnologías móviles. *Revista de Comunicación y Salud, 10*(1), 19-41. http://doi.org/10.35669/rcys.2020.10(1).19-41