CUADERNOS.INFO Nº 49
Versión electrónica: ISSN 0719-367x
http://www.cuadernos.info
https://doi.org/10.7764/cdi.49.27817

BY-SA

Recibido: 26-11-2020 / Aceptado: 19-04-2021

Detectando el odio ideológico en Twitter. Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español

Detecting ideological hatred on Twitter. Development and evaluation of a political ideology hate speech detector in tweets in Spanish

Detectando o ódio ideológico no Twitter. Desenvolvimento e avaliação de um detector de discurso de ódio por ideologia política no Twitter em espanhol

Javier J. Amores, Universidad de Salamanca, Salamanca, España (javieramores@usal.es)

David Blanco-Herrero, Universidad de Salamanca, Salamanca, España (david.blanco.herrero@usal.es)

Patricia Sánchez-Holgado, Universidad de Salamanca, Salamanca, España (patriciasanc@usal.es)

Maximiliano Frías-Vázquez, Universidad de Salamanca, Salamanca, España (maxfrias@usal.es)

RESUMEN | El discurso de odio propagado a través de redes sociales como Twitter merece atención especial, ya que su incremento puede relacionarse con el aumento de crímenes de odio. De las 11 categorías de discriminación que contempla el Ministerio de Interior de España, la segunda en la que más delitos de odio se registran al año es la ideología. Sin embargo, esta categoría queda fuera de la mayor parte de los planes de acción para estudiar y combatir los delitos de odio. Lo mismo ocurre con los trabajos académicos, que se centran mayoritariamente en el odio en inglés y a nivel general. Los que estudian un único tipo de odio se han enfocado en el racismo, la xenofobia o la discriminación de género, pero nunca en la ideología política. Asimismo, los prototipos de detección desarrollados hasta ahora no usan bases de datos generadas manualmente por varios codificadores. Esta investigación busca

FORMA DE CITAR

Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P. & Frías-Vázquez, M. (2021). Detectando el odio ideológico en Twitter. Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos.info, (49)*, 98-124. https://doi.org/10.7764/cdi.49.27817

superar estas limitaciones, desarrollando y evaluando un detector automático de discurso de odio por motivos ideológicos en Twitter en español a partir de técnicas de aprendizaje automático supervisado. Para ello, se ha desarrollado un total de ocho modelos predictivos a partir de un corpus de entrenamiento generado ad-hoc, y haciendo uso de modelado superficial y de aprendizaje profundo, lo que permite mejorar el rendimiento final del prototipo. El desarrollo del corpus permitió observar, además, que un 16,2% de la muestra, recogida en el otoño de 2019, incluyó algún tipo de odio ideológico.

PALABRAS CLAVE: discurso de odio; odio en línea; Twitter; ideología política; aprendizaje profundo; aprendizaje automático; clasificación supervisada.

ABSTRACT | *Hate speech spread through social media such as Twitter deserves special attention, as its increase may be related to the rise in hate crimes. Of the 11 categories of discrimination contemplated by the Spanish Ministry of Internal Affairs, the second in which the most hate crimes are registered per year is political ideology. However, this category falls outside of most action plans to study and combat hate crimes. The same occurs in the case of academic works since most focus on analyzing and detecting hate in English and at a general level. The few authors who have targeted their studies to a single type of hate to improve accuracy, have focused on racism, xenophobia, or gender discrimination, but never on political ideology. Furthermore, the detection prototypes developed so far have not used databases generated manually by various coders. This paper aims to overcome these limitations, developing and evaluating an automatic hate speech detector on Twitter in Spanish for reasons of ideological discrimination, using supervised machine learning techniques. For this, we developed a total of eight predictive models from an ad-hoc generated training corpus, and making use of shallow modelling, but also deep learning, which has allowed to improve the final performance of the prototype. In addition, the development of the corpus allowed us to observe that 16.2% of the sample, collected in autumn 2019 and manually analyzed, included some type of ideological hatred.*

KEYWORDS: *hate speech; online hate; Twitter; political ideology; deep learning; machine learning; supervised classification.*

RESUMO | O discurso de ódio que se espalha pelas redes sociais como o Twitter merece atenção especial, pois seu aumento pode estar relacionado ao aumento dos crimes de ódio. Das onze categorias de discriminação contempladas pelo Ministério do Interior da Espanha, a segunda em que mais crimes de ódio são registrados por ano é a ideologia política. No entanto, esta categoria está fora da maioria dos planos de ação para estudar e combater os crimes de ódio. O mesmo acontece com os trabalhos acadêmicos, já que a maioria concentra-se em analisar e detectar o ódio em inglês e

em um nível geral, e os poucos autores que limitaram seus estudos a um único tipo de ódio concentraram-se no racismo, xenofobia ou discriminação de gênero, mas nunca na ideologia política. Além disso, os protótipos de detecção desenvolvidos até o momento não usaram bancos de dados gerados manualmente por vários codificadores. A presente pesquisa visa superar essas limitações, desenvolvendo e avaliando um detector automático de discurso de ódio no Twitter em espanhol por motivos de discriminação ideológica, baseado em técnicas de aprendizagem automática supervisionada. Para isso, foram desenvolvidos um total de 8 modelos preditivos a partir de um corpus de treinamento gerado ad-hoc, e fazendo uso de modelagem superficial, mas também de aprendizagem profunda, que tem permitido melhorar o desempenho final do protótipo. O processo de elaboração do corpus também nos permitiu observar que 16,2% da amostra, coletada no outono de 2019, incluía algum tipo de ódio ideológico.

PALAVRAS-CHAVE: discurso de ódio; ódio online; Twitter; ideologia política; aprendizagem profunda; aprendizagem de máquina; classificação supervisionada.

INTRODUCCIÓN Y JUSTIFICACIÓN

El discurso de odio merece una especial atención académica debido a sus implicaciones sociales, ya que puede ser un importante precursor de delitos más graves, que se han incrementado en los últimos años (Organization for Security and Cooperation in Europe (OSCE), 2020). En este sentido, Müller y Schwarz (2020) plantean que existe una correlación entre el aumento del odio en línea y los crímenes de odio cometidos en determinadas regiones y contextos, por lo que es fundamental estudiar este tipo de mensajes que se transmiten en la red con el fin de prevenir y contrarrestar sus efectos.

El odio parece haber encontrado en las plataformas sociales el entorno ideal para propagarse, especialmente en Twitter por su papel en la formación de opinión pública, gracias a su volumen de uso, un 16% de la población española según el Reuters Institute Digital News Report (Newman et al., 2019), y a la presencia de políticos y periodistas (Rodríguez & Ureña, 2011). Esta capacidad de las redes sociales para marcar la agenda pública también se sustenta en el interés que reciben desde los medios tradicionales (Bane, 2019).

En esta red social, los mensajes que expresan odio, rechazo, intolerancia o discriminación hacia ciertos grupos vulnerables no han dejado de aumentar en los últimos años, propagados por todo tipo de usuarios. Incluso durante la reciente crisis sanitaria, el discurso de odio en Twitter ha seguido aumentando, hacia enfermos, mayores, migrantes, extranjeros, pero también hacia los dirigentes políticos, que han ido extremando sus discursos. Este incremento del odio se observa en los últimos informes *Online Hate and Harassment* de la Anti-Defamation League (2020, 2021), que reflejan un aumento exponencial de todas las formas de ciberodio en la mayor parte de las redes sociales desde 2018. Recientes estudios han evidenciado una tendencia negativa en la representación de los migrantes y refugiados que transmiten los principales medios de los países del Mediterráneo (Amores et al., 2020) y de Europa occidental (Amores et al., 2019), lo que se puede relacionar directamente con el incremento del odio por racismo o xenofobia, e indirectamente con la ideología.

Esta situación se percibe especialmente en España, uno de los países en los que las recientes crisis han tenido efectos más visibles, lo que puede haber potenciado los discursos extremistas y radicales (Ferreira, 2019). Además, del aumento de los discursos de odio en línea y *offline* por razones ideológicas, España suma la ausencia de una estrategia nacional independiente y dirigida a prevenir este tipo de delitos. Si bien en septiembre de 2018 el gobierno firmó un acuerdo de cooperación institucional con el Consejo General del Poder Judicial y la Fiscalía General del Estado para luchar contra la intolerancia (Ministerio de Empleo, Migraciones y

Seguridad Social, 2018), en este no se define un plan de acción articulado para combatir el incremento de los delitos de odio y, específicamente, del discurso de odio, que suele constituir la raíz del resto de crímenes. Esto hace más necesario implementar nuevos métodos que ayuden a identificar y monitorizar de manera automática y a gran escala el discurso de odio en línea para poder prevenirlo y combatirlo, contrarrestando al mismo tiempo el odio en las calles, y con ello, los crímenes de odio. En un entorno digital libre de vigilancia y regulación, existe una creciente preocupación por las potenciales víctimas de este odio en línea, algo que evidencia el último Informe Raxen (Movimiento contra la Intolerancia, 2019). En este sentido, es difícil justificar el volumen de odio existente en una red social de manera precisa, dada su fluctuación temporal en función de eventos mediáticos (Arcila Calderón et al., 2020); sin embargo, sí existe una tendencia creciente en el volumen de crímenes de odio, tanto en España como en el resto de Europa (OSCE, 2020), y dada la conexión existente entre el odio en línea y estos delitos (Müller & Schwarz, 2020), la relevancia del estudio del ciberodio es incuestionable.

La ideología política es la segunda categoría de discriminación por volumen de delitos de odio después del racismo y la xenofobia, según el Informe sobre la Evolución de los Delitos de Odio en España (Ministerio de Interior, 2020). Por esta razón, el objetivo de este trabajo es desarrollar y evaluar un detector automático del discurso de odio por razones ideológicas propagado a través de Twitter en español, haciendo uso de métodos computacionales. Para ello, se parte de una muestra previamente seleccionada con un filtro de palabras clave temáticas identificadas manualmente a lo largo del otoño de 2019.

DEFINIENDO EL DISCURSO DE ODIO EN LÍNEA

El discurso de odio no es una preocupación exclusiva de las sociedades actuales, sino que tradicionalmente ha existido como una forma radical de expresar el rechazo y la intolerancia frente a la otredad (Krippendorf, 2010). Ya en 1997, Calvert señalaba este tipo de discurso como una problemática a analizar, comprender y combatir con enfoques comunicacionales, involucrando necesariamente a todos los elementos de los modelos de transmisión de la comunicación (fuente, mensaje, canal y receptor). No obstante, este tipo de discurso preocupa especialmente en la actualidad debido al auge de las redes sociales y al nuevo perfil de los prosumidores (Carmona, 2010), con seguidores para propagar contenidos sin regulación, además de la demostrada influencia social y sobre la opinión pública (Isasi & Juanatey, 2017). Por eso, se contemplan como un posible delito a investigar y se debate la necesidad, urgencia y dificultad de su detección y eliminación (Jubany & Roiha, 2018; Tamarit Sumalla, 2018).

Antes de afrontar cualquier estrategia de detección del discurso de odio en los entornos digitales, es conveniente tratar de definirlo. En este sentido, aunque aún no exista una conceptualización única y estandarizada del discurso de odio por la propia amplitud y subjetividad del término, varios autores han propuesto una definición y taxonomía, discutiendo acerca de los tipos y niveles de discursos odiosos que se dan en la actualidad, con base en si podrían ser considerados como delito, o concebir dentro de los márgenes de la libertad de expresión. En esta línea, Benesch (2014) se aleja de la terminología de odio para proponer el término discurso peligroso (*dangerous speech*), con el que se refiere a aquellos discursos que tienen una considerable probabilidad de desencadenar episodios de violencia. Leader Maynard y Benesch (2016) sostienen que tanto este discurso como la ideología peligrosa que lo fomenta constituyen un riesgo real de terminar convirtiéndose en crímenes y atentados, por lo que es necesario monitorizar y combatir toda expresión de odio dada su peligrosidad. Por su parte, Gagliardone y sus colegas (2015) entienden como discurso de odio todo tipo de expresiones que inciten directamente a la comisión de actos de discriminación o violencia por motivos de odio racial, xenófobo, por orientación sexual u otras formas de intolerancia, extendiendo además el término a aquellas expresiones que fomentan el prejuicio, considerando que pueden contribuir indirectamente a que se genere un clima de hostilidad que pueda llegar a propiciar actos discriminatorios o ataques violentos. Según estos autores, en la actualidad se ha generalizado el uso del término discurso de odio para referirse a un conglomerado heterogéneo de manifestaciones que engloba desde amenazas a individuos o colectivos hasta casos en los que algunas personas simplemente expresan su ira contra las autoridades de manera más o menos ofensiva. Sin embargo, el conflicto reside en los difusos límites de lo conceptualizado como discurso de odio, que a menudo puede entrar en conflicto con el derecho fundamental a la libertad de expresión, y en cómo discernir qué parte de esta compleja amalgama de discursos puede constituir un delito. Algo que, como explica Arroyo (2017), suele residir en la mera interpretación teórica y jurisprudencial del código penal, que en el caso español tan solo se refiere a este tipo de delitos en el artículo 510 de la Ley Orgánica 10/1995, relativo a la incitación al odio.

Tratando de ayudar a resolver este conflicto, Miró Llinares (2016) ofrece, además de una amplia definición, una taxonomía que permite diferenciar entre el tipo de discurso de odio que pudiera constituir un delito por ser más explícito, directo o instigar a la violencia física, y aquel más sutil que, aunque suponga una ofensa y exprese rechazo hacia ciertos individuos o grupos vulnerables, puede enmarcarse en los márgenes de la libertad de expresión. No obstante, a la hora de monitorizar y combatir el ciberodio conviene considerar todos los niveles en

los que se puede representar y propagar, ya que, por un efecto acumulativo, todos pueden contribuir de la misma manera a la hora de generar deshumanización, estigmatización y, en última instancia, episodios de violencia hacia cualquier tipo de otredad (Isasi & Juanatey, 2017).

Por otro lado, a nivel institucional, la Unión Europea ha tratado de definir los límites de la libertad de expresión, acotando cada vez más la conceptualización del discurso de odio, aunque sin mucho éxito práctico por no contar con un reflejo patente en la jurisprudencia de los distintos países miembros. Para el Consejo de Europa, a través de su Recomendación No. R (97)20 del Comité de Ministros sobre discurso de odio (Council of Europe, 1997), este discurso es entendido como la promoción de mensajes que impliquen el rechazo, el menosprecio, la humillación, el acoso, el descrédito y la estigmatización de individuos o colectivos sociales basados en unos atributos particulares. Siendo así, para que un discurso pueda ser considerado un delito de odio, debe propagar, incitar, promover o justificar el odio racial, la xenofobia, el antisemitismo y otras formas de odio basadas en la intolerancia. En esta línea, la Comisión Europea contra el Racismo y la Intolerancia, mediante su Recomendación General nº15 sobre Combatir el Discurso de Odio (European Commission against Racism and Intolerance, 2016), especifica que el odio puede venir motivado por razones de raza, color, ascendencia, origen nacional o étnico, ideología, edad, discapacidad, lengua, religión, sexo, género, identidad de género, orientación sexual y otras características o condiciones personales. El Ministerio de Interior de España (2020), en su último Informe de Evaluación sobre delitos de Odio en el país, recoge un total de 11 categorías de discriminación en las que se pueden clasificar los delitos cometidos hacia públicos vulnerables. Estos son: (1) racismo/xenofobia, (2) ideología política, (3) orientación sexual e identidad de género, (4) creencias o prácticas religiosas, (5) discapacidad, (6) razones de género, (7) antisemitismo, (8) aporofobia, (9) antigitanismo, (10) discriminación generacional, y (11) discriminación por enfermedad. Las tres primeras son las que motivan un mayor número de delitos de odio en España al año, según las cifras recogidas por los últimos informes del Ministerio, siendo la segunda –la ideología política–, el tipo de discriminación que más se ha incrementado en los últimos años, especialmente en los espacios digitales. Sin embargo, esta suele quedar fuera de los márgenes de interés social, institucional y académico al estudiar y analizar el discurso de odio. Contando con estas premisas, este trabajo se enfoca en detectar el discurso de odio motivado específicamente por razones de ideología política. También se trata de abarcar todos los niveles de odio tipificados, procurando ampliar la detección del discurso de odio en Twitter, dado que se espera que el más explícito y que pudiera considerarse delito no tenga una presencia abultada en el contexto español.

DETECTAR EL DISCURSO DE ODIO EN LÍNEA POR RAZONES DE IDEOLOGÍA POLÍTICA

En los últimos años, numerosos autores han estudiado estos discursos desde diversas perspectivas. Chetty y Alathur (2018) lo analizan desde la base jurisprudencial, concluyendo que las medidas políticas adecuadas, así como las actuaciones de las propias redes sociales, son esenciales para contrarrestar el discurso de odio de manera efectiva. Otros, como ElSherief y sus colegas (2018), lo estudian usando una perspectiva lingüística y psicolingüística basada en datos, ofreciendo un marco de entendimiento desde el cual poder identificar el odio que se transmite en las redes sociales. Con un enfoque de detección más automatizado y masivo, Mondal y sus colegas (2017) proponen un sistema para medir y monitorear el discurso de odio propagado en las redes sociales Twitter y Whisper a partir de expresiones y palabras clave determinadas, y centrando la atención en reconocer los principales objetivos a los que se dirige el odio de manera masiva. Malmasi y Zampieri (2017), por su parte, proponen un método de detección del odio propagado en redes sociales basándose en procesamiento del lenguaje natural y técnicas de clasificación supervisada.

Estos trabajos se centran en el discurso de odio en línea como un problema a detectar y combatir, y lo tratan desde un punto de vista genérico e internacional, es decir, tratando de identificar el discurso de odio propagado en inglés, motivado por todo tipo de razones, dirigido hacia todo tipo de públicos y en cualquier momento y lugar, lo que supone un enfoque muy ambicioso que podría presentar problemas de validez interna, especialmente en las estrategias a gran escala. Incluso los prototipos desarrollados recientemente por Salminen et al. (2020), de los más innovadores y avanzados porque usan aprendizaje profundo e incluyen la detección en diversas fuentes en línea, recaen en ese enfoque genérico. Esto lo convierte en una limitación, debido a que los modelos resultantes no logran ser tan eficaces, fiables y, paradójicamente, generalizables como los que son entrenados con ejemplos reales de un solo tipo de odio y en una categoría discriminatoria específica, diferenciando por tanto conceptos, características y matices lingüísticos.

En este sentido, cabe señalar que en el panorama internacional sí existe algún ejemplo de estrategia de detección del ciberodio que tiene en cuenta niveles, categorías de prejuicio, o los grupos vulnerables víctimas de ese discurso. Destacan trabajos como el de Davidson y sus colegas (2017), que diferencia entre mensajes de odio directo y mensajes ofensivos, o el desarrollado por Badjatiya y sus colegas (2017), que identifica mensajes con contenidos racistas o sexistas usando modelado profundo. Asimismo, la mayoría de los trabajos citados tienen en común una segunda limitación, y es que no han usado corpus de entrenamiento generados ad-hoc. La mayor parte de los prototipos existentes hasta el momento basan la detección en diccionarios lexicon previamente desarrollados, o, en el caso de usar corpus de

ejemplos para entrenar los algoritmos de clasificación, suelen emplear bases de datos ya disponibles de otros autores y trabajos anteriores, como ocurre con el desarrollado por Salminen et al. (2020), lo que también influye en la validez interna del prototipo y en su fiabilidad final. En el contexto españolde los pocos trabajos que pretenden abordar el problema de la detección del discurso de odio en línea en español son los de Pereira Kohatsu (2017) y de Pereira Kohatsu y sus colegas (2019). Este prototipo presenta las mismas limitaciones que la mayor parte de los internacionales expuestos previamente, ya que, aunque desarrolló un corpus de entrenamiento ad-hoc para generar los modelos predictivos, fue elaborado por un único codificador, lo que supone un problema de validez interna por su potencial subjetividad.

Por todas estas razones, este trabajo se centra en desarrollar un prototipo capaz de detectar el discurso de odio propagado en Twitter en español por razones de ideología política. Hasta ahora, Arcila Calderón, Valdez Apolo, Blanco Herrero y Amores son de los pocos autores que han centrado su atención en analizar y detectar el rechazo manifestado en Twitter por motivos de discriminación concretos, en específico el racismo y la xenofobia. Para ello, usaron en primer lugar métodos de análisis manuales (Valdez-Apolo et al., 2019) para posteriormente desarrollar una técnica de detección automática y a gran escala basada en aprendizaje automático supervisado y haciendo uso de los corpus elaborados previamente de manera manual (Arcila-Calderón et al., 2020). Siguiendo esta línea, el objetivo de este trabajo es desarrollar una estrategia de detección más avanzada y centrada en el odio por ideología política. En este sentido, cabe señalar que los mensajes de carácter político transmitidos a través de Twitter han sido analizados en numerosas ocasiones, incluso en el contexto español, pero normalmente con la finalidad de estudiar el uso que hacen de esta plataforma social los propios partidos o líderes políticos (Marín Dueñas & Díaz Guerra, 2016; López-García, 2016), de analizar los contextos que rodean a las campañas y jornadas electorales (López-Meri, 2017; García-Ortega & Zugasti-Azagra, 2018), o de detectar la orientación ideológica y predecir resultados electorales (Alonso González, 2017; Said-Hung et al., 2017). Arcila Calderón y sus colegas (2017) desarrollaron previamente una estrategia de detección de los sentimientos políticos en Twitter en español basada en clasificación supervisada, y que también podría ser aplicada al análisis de los contextos políticos, al análisis del apoyo a los distintos partidos y a la predicción de resultados electorales. Sin embargo, hasta ahora ningún trabajo ha centrado su atención en el análisis y detección del discurso de odio por razones políticas.

Con estas premisas, se pretende solventar y superar las limitaciones reseñadas a partir de una serie de elementos diferenciadores. Primero, usar técnicas de aprendizaje automático supervisado para generar bases de datos propias, elaboradas ad-hoc con ejemplos clasificados manualmente y con total acuerdo inter-jueces, que

sirvan como corpus de entrenamiento para los modelos predictivos resultantes. En segundo lugar, la elaboración de un corpus de entrenamiento específico de ideología política, para generar unos modelos predictivos más fiables. En este sentido, ya que la creación de los corpus de entrenamiento requiere la clasificación manual de ejemplos previamente descargados y filtrados de las API de Twitter, se plantea la siguiente pregunta de investigación: ¿Qué frecuencia y porcentaje de tuits de odio por ideología política se detectan a través de la clasificación manual en una muestra de tuits previamente filtrados? (PI1).

El tercer elemento innovador es el uso de aprendizaje profundo para generar los modelos predictivos. Específicamente, este trabajo usa redes neuronales recurrentes, un algoritmo que, a priori, debería presentar ventajas significativas frente a los algoritmos de clasificación tradicionales, ofreciendo un mejor rendimiento especialmente aplicado a clasificar textos. En este sentido, se plantean las siguientes preguntas: ¿Qué algoritmo de aprendizaje automático presenta mejor rendimiento para generar un modelo predictivo capaz de detectar el discurso de odio por ideología política en Twitter en español? (PI2); ¿Presenta mejor rendimiento el aprendizaje profundo que el aprendizaje superficial para generar modelos capaces de detectar el discurso de odio por ideología política en Twitter en español? (PI2A).

MÉTODO

Para desarrollar el detector de discurso de odio por ideología política en Twitter se siguió una estrategia de detección a gran escala basada en el cómputo intensivo de datos bajo la infraestructura de supercomputación de Castilla y León, Scayle, y haciendo uso de técnicas de procesamiento del lenguaje natural y de aprendizaje automático supervisado. Para ello, el trabajo se dividió en tres fases principales, esquematizadas en la figura 1.

Fase exploratoria

En esta primera fase se llevó a cabo una exploración cualitativa en profundidad del discurso de odio por razones de ideología que se propaga en medios sociales como Twitter. También se hizo una revisión de literatura como aproximación teórica y se identificaron cuentas, perfiles y hashtags a través de los que se publica una mayor cantidad de mensajes de odio por ideología política. La exploración de esas fuentes potenciales de odio en Twitter serviría para entender y acotar las diferentes formas en las que se expresa el odio motivado por la ideología, los distintos contextos en los que se propaga, así como los términos y expresiones más usadas. Esto ayudó a generar más tarde los filtros lingüísticos que permitirían descargar los tuits para su clasificación manual.

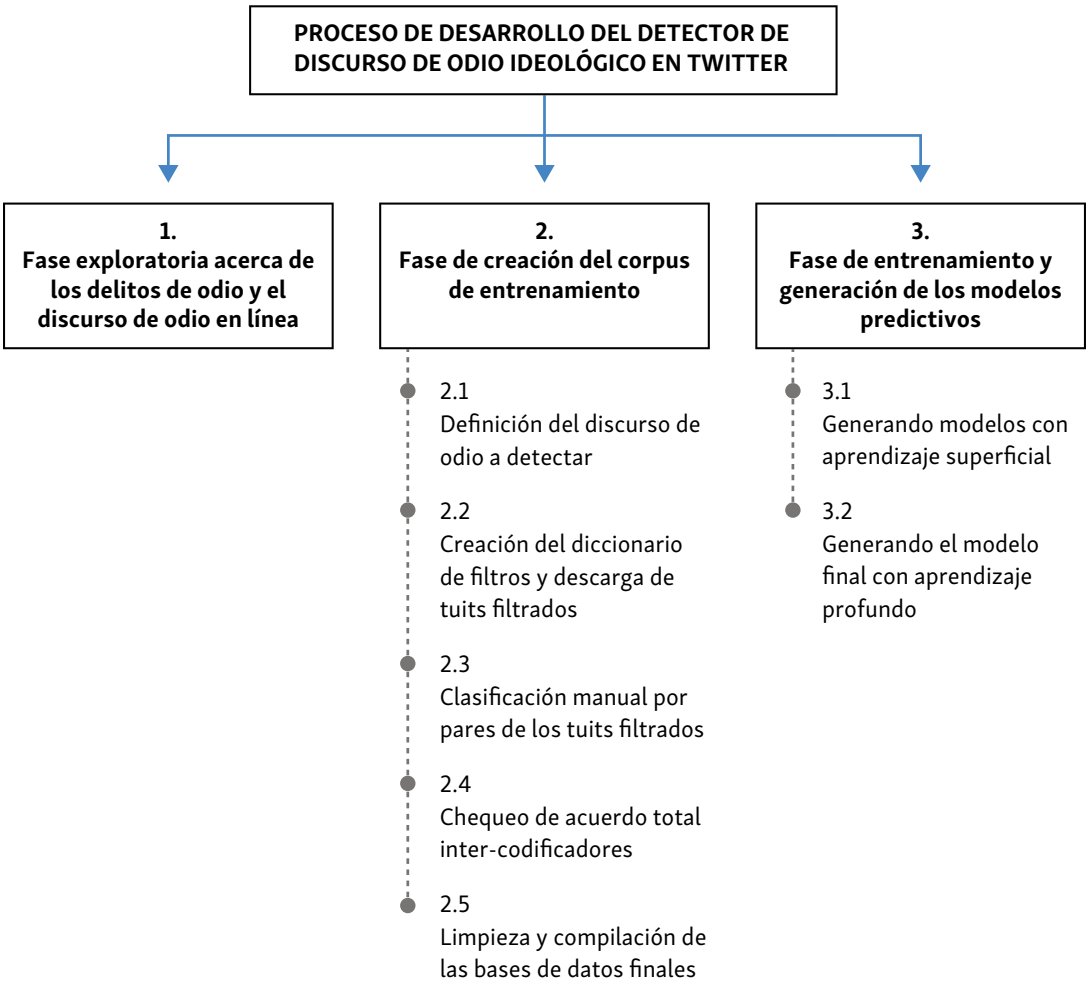

Figura 1. Proceso metodológico llevado a cabo para desarrollar el Detector automático de discurso de odio por ideología política en Twitter en español

Fuente: Elaboración propia.

Fase de creación del corpus de entrenamiento

En esta fase se procedió a crear las bases de datos ad-hoc a partir de ejemplos fiables del tipo de odio a detectar. Estas bases servirían como corpus para entrenar los modelos predictivos que finalmente permitirían detectar los mensajes de odio de manera automática y masiva. Es la etapa más larga y laboriosa, que permite superar las limitaciones de los prototipos desarrollados anteriormente, que usaban diccionarios o bases de datos generales y preexistentes. La integran una serie de subetapas, descritas a continuación.

Definición y tipología del discurso de odio a detectar

En primer lugar, se establecieron criterios que acotaran el tipo de discurso a detectar para poder generar las bases de datos a medida. Según las posibilidades

identificadas en la exploración cualitativa, y considerando tanto las definiciones aportadas por los distintos autores e instituciones como el propio marco legal europeo, se amplió la definición de discurso de odio, abarcando las distintas acepciones y tipos que se ofrecen desde la academia, las instituciones públicas y el código penal español, así como los tres niveles de odio en línea aportados por Miró Llinares (2016). Así, para generar las bases de datos se incluyeron todos los tipos de discursos odiosos que pueden constituir un delito, pero también aquellos más sutiles que, a priori, pudieran considerarse dentro del abanico de la libertad de expresión. Esto se determinó debido a que, en la fase previa, se había detectado una parte muy minoritaria de odio directo y explícito, y la intención era poder detectar la mayor parte posible de mensajes con este tipo de contenidos. Además, en el proceso de validación de la clasificación manual llevada a cabo siguiendo las bases de un análisis de contenido, y en el posterior entrenamiento de los modelos, los resultados se refinarían, quedando los ejemplos más seguros, filtrando y rechazando los dudosos o ambiguos por no tener acuerdo intercodificadores, motivo por el que también interesaba abarcar el mayor abanico posible de formas de odio. Así, dada la escasez de mensajes con discurso de odio ilegal en Twitter dentro del marco español, se decidió entrenar a los modelos para detectar todo nivel de contenidos odiosos. También se definió lo que se consideraría discurso de odio por ideología política, recopilando todos los términos despectivos, expresiones y objetivos recogidos en la fase exploratoria.

Creación de diccionario de filtros y descarga de tuits

Una vez definidos los niveles y tipos de discurso de odio, se generó un diccionario de palabras y sus combinaciones que pudieran servir de filtro para descargar inicialmente potenciales tuits con odio por ideología política. Para ello, se usaron las cuentas y etiquetas de Twitter propagadoras de una mayor cantidad de discurso de odio en España por razones ideológicas. Posteriormente, estos mensajes se clasificaron manualmente por públicos referenciados y por inclusión de odio.

En segundo lugar, basándose en estos ejemplos, se realizó una selección final de términos de búsqueda, en un formato de lista de palabras, raíces o combinaciones de palabras que pudieran ser representativas de odio por ideología política, siguiendo la distinción realizada por Kalampokis y sus colegas (2013) para conformar el diccionario definitivo que serviría de filtro para la descarga. A continuación, se tradujo al lenguaje computacional (figura 2) para poder descargar la cantidad necesaria de tuits desde las API de Twitter. De esta forma, aunque se descargó una mayor cantidad de mensajes, finalmente se recolectó y compiló una muestra de 24.000 tuits en una base de datos para su posterior clasificación manual.

```
word = ['podemita', 'trifachito', 'falangito', 'rojo\ncomunista', 'extrema\nizquierda', 'extrema\nderecha', 'izmierda',
    'independentista\nfacha', 'independentista\nfascista', 'independentista\ngentuza', 'independentista\nbasura',
'independentista\nfacha','independentista\nimbecil', u'independentista\nimbécil', 'independentista\ngilipollass',
'independentista\nlacra', 'independentista\nescoria', 'independentista\asco', 'independentista\nmierda',
'independentista\nnazi', 'independentista\nputo', 'independentista\nputa', 'independentista\nmaldito',
'independentista\nmaldita', 'independentista\nsucio', 'independentista\nsucia',
    'independentismo\nfacha', 'independentismo\nfascista', 'independentismo\ngentuza', 'independentismo\nbasura',
'independentismo\nfacha','independentismo\nimbecil', u'independentismo\nimbécil', 'independentismo\ngilipollas',
'independentismo\nlacra', 'independentismo\nescoria', 'independentismo\asco', 'independentismo\nmierda',
'independentismo\nnazi', 'independentismo\nputo', 'independentismo\nputa', 'independentismo\nmaldito',
'independentismo\nmaldita', 'independentismo\nsucio', 'independentismo\nsucia',
    'socialista\nfacha', 'socialista\nfascista', 'socialista\ngentuza', 'socialista\nbasura',
'socialista\nfacha','socialista\nimbecil', 'socialista\ngilipollas', 'socialista\nlacra', 'socialista\nescoria', 'socialista\asco',
'socialista\nmierda', 'socialista\nnazi', 'socialista\nputo', 'socialista\nputa', 'socialista\nmaldito', 'socialista\nmaldita',
'socialista\nsucio', 'socialista\nsucia',
    'socialismo\nfacha', 'socialismo\nfascista', 'socialismo\ngentuza', 'socialismo\nbasura',
'socialismo\nfacha','socialismo\nimbecil', u'socialismo\nimbécil', 'socialismo\ngilipollass', 'socialismo\nlacra',
'socialismo\nescoria', 'socialismo\asco', 'socialismo\nmierda', 'socialismo\nnazi', 'socialismo\nputo', 'socialismo\nputa',
'socialismo\nmaldito', 'socialismo\nmaldita', 'socialismo\nsucio', 'socialismo\nsucia',
    'nacionalista\nfacha', 'nacionalista\nfascista', 'nacionalista\ngentuza', 'nacionalista\nbasura',
'nacionalista\nfacha','nacionalista\nimbecil', 'nacionalista\ngilipollas', 'nacionalista\nlacra', 'nacionalista\nescoria',
'nacionalista\asco', 'nacionalista\nmierda', 'nacionalista\nnazi', 'nacionalista\nputo', 'nacionalista\nputa',
'nacionalista\nmaldito', 'nacionalista\nmaldita', 'nacionalista\nsucio', 'nacionalista\nsucia',
    'nacionalismo\nfacha', 'nacionalismo\nfascista', 'nacionalismo\ngentuza', 'nacionalismo\nbasura',
'nacionalismo\nfacha','nacionalismo\nimbecil', 'nacionalismo\ngilipollas', 'nacionalismo\nlacra', 'nacionalismo\nescoria',
'nacionalismo\asco', 'nacionalismo\nmierda', 'nacionalismo\nnazi', 'nacionalismo\nputo', 'nacionalismo\nputa',
'nacionalismo\nmaldito', 'nacionalismo\nmaldita', 'nacionalismo\nsucio', 'nacionalismo\nsucia',
    'comunista\nfacha', 'comunista\nfascista', 'comunista\ngentuza', 'comunista\nbasura',
'comunista\nfacha','comunista\nimbecil', 'comunista\ngilipollas', 'comunista\nlacra', 'comunista\nescoria',
'comunista\asco', 'comunista\nmierda', 'comunista\nnazi', 'comunista\nputo', 'comunista\nputa', 'comunista\nmaldito',
'comunista\nmaldita', 'comunista\nsucio', 'comunista\nsucia',
    'comunismo\nfacha', 'comunismo\nfascista', 'comunismo\ngentuza', 'comunismo\nbasura',
'comunismo\nfacha','comunismo\nimbecil', 'comunismo\ngilipollas', 'comunismo\nlacra', 'comunismo\nescoria',
'comunismo\asco', 'comunismo\nmierda', 'comunismo\nnazi', 'comunismo\nputo', 'comunismo\nputa',
'comunismo\nmaldito', 'comunismo\nmaldita', 'comunismo\nsucio', 'comunismo\nsucia',
    'golpista\nfacha', 'golpista\nfascista', 'golpista\ngentuza', 'golpista\nbasura', 'golpista\nfacha','golpista\nimbecil',
'golpista\ngilipollas', 'golpista\nlacra', 'golpista\nescoria', 'golpista\asco', 'golpista\nmierda', 'golpista\nnazi',
'golpista\nputo', 'golpista\nputa', 'golpista\nmaldito', 'golpista\nmaldita', 'golpista\nsucio', 'golpista\nsucia',
    'golpismo\nfacha', 'golpismo\nfascista', 'golpismo\ngentuza', 'golpismo\nbasura',
'golpismo\nfacha','golpismo\nimbecil', 'golpismo\ngilipollas', 'golpismo\nlacra', 'golpismo\nescoria', 'golpismo\asco',
'golpismo\nmierda', 'golpismo\nnazi', 'golpismo\nputo', 'golpismo\nputa', 'golpismo\nmaldito', 'golpismo\nmaldita',
'golpismo\nsucio', 'golpismo\nsucia',
```

Figura 2. Fragmento del script final empleado para la descarga filtrada de potenciales tuits de odio por ideología política

Fuente: Elaboración propia.

Clasificación manual por pares

Posteriormente, se procedió a clasificar manualmente los mensajes a través de la plataforma Doccano, que facilitó la tarea de etiquetado de los textos entre varios codificadores (figura 3). Así, todos los tuits fueron clasificados por un codificador principal y ocho secundarios (3000 tuits cada uno). Para posteriormente cruzar los resultados y que los mensajes resultantes fueran más fiables, los jueces secundarios debían ser personas externas y ajenas al proyecto, por lo que se eligió a estudiantes de grado y posgrado de la Universidad de Salamanca, quienes fueron entrenados previo a la clasificación y a los que se les entregó el manual elaborado con ejemplos a modo de libro de códigos. Se etiquetaron los mensajes de manera binaria en odio y no odio, al mismo tiempo que el codificador principal descartaba mensajes desvinculados del tema trabajado.

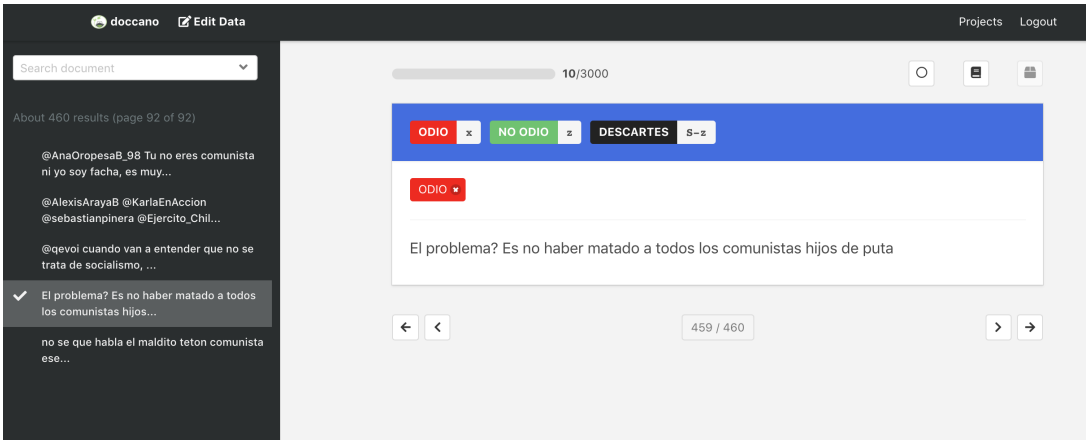

Figura 3. Clasificación manual en la plataforma Doccano

Fuente: Elaboración propia.

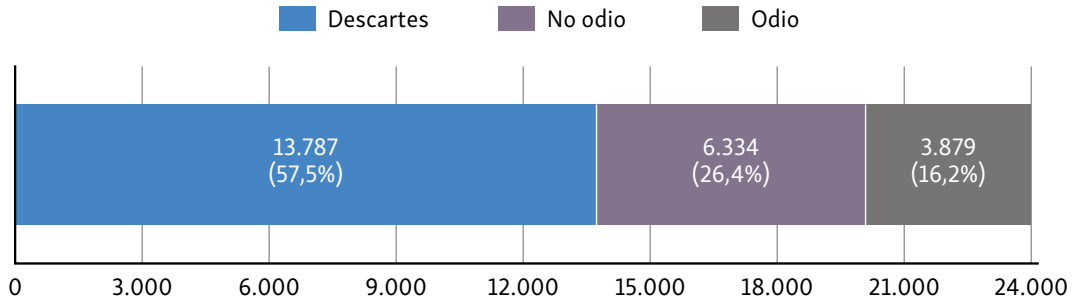

Figura 4. Frecuencias y porcentajes de la clasificación manual de tuits políticos

Fuente: Elaboración propia.

Chequeo de acuerdo total inter-codificadores

Una vez clasificados todos los tuits por dos codificadores, se revisó la fiabilidad inter-codificadores, manteniendo solo aquellos mensajes que fueron clasificados con la misma etiqueta por ambos codificadores y desechando el resto. Es decir, la fiabilidad inter-codificadores sería de α=1. Este paso, además de garantizar la calidad de la codificación permite superar una de las principales limitaciones de algunos prototipos como el de Pereira-Kohatsu (2017), mencionado anteriormente.

Limpieza y compilación de las bases de datos finales

Finalizado ese proceso, se procedió a limpiar las bases de datos, con lo que el corpus de entrenamiento resultó en 16,2% tuits de odio fiables (N=3879) y 26,4% de no odio (N=6334) (figura 4).

Generando los modelos predictivos

Contando con el corpus de entrenamiento validado, este se utilizó para entrenar y generar los modelos predictivos que finalmente permitirían detectar el discurso de odio en Twitter en español por razones de ideología política de manera automática y a gran escala. Se generaron un total de ocho modelos predictivos: seis usando algoritmos de aprendizaje superficial, uno generado a partir de los votos de esos modelos anteriores y otro usando aprendizaje profundo.

Modelado superficial

Los seis modelos predictivos que se generaron haciendo uso de algoritmos de clasificación tradicionales se basaron en *Bag of Words* como representación del texto, a partir del cual cada palabra es tomada como un vector. Se utilizaron las librerías NLTK y SciKit-Learn de Python para generar modelos de clasificación binaria y los siguientes algoritmos convencionales de aprendizaje superficial: Naïve Bayes Original, Naïve Bayes Multinomial, Naïve Bayes para modelos multivariados Bernoulli, Regresión Logística, Regresión Lineal con Gradiente Descendiente Estocástico y Máquinas de Vectores de Soporte. También se aplicaron técnicas de procesamiento de lenguaje natural (NLP) para extraer las características del conjunto de mensajes etiquetados. En el proceso de entrenamiento de los modelos, las palabras más repetidas del conjunto de ejemplos que formaban el corpus de entrenamiento fueron *tokenizadas* y convertidas en características cuantitativas o vectores con los que los modelos predictivos pudieran trabajar. En este proceso de modelado, cada uno de los corpus fue dividido aleatoriamente en dos subgrupos: 70% dedicado al entrenamiento y 30% dedicado a prueba y validación de los propios modelos. Así, se generaron clasificadores optimizados para cada uno de los seis algoritmos mencionados y se implementaron sobre el corpus de entrenamiento para generar seis modelos predictivos capaces de detectar el discurso de odio en tuits en español por razones de ideología política. Una vez desarrollados estos modelos, se generó un modelo basado en el voto de cada uno de los seis anteriores. Este clasificador elige la categoría –odio/no odio– que la mayoría de los modelos predice (en el supuesto de empate, lo hace aleatoriamente), añadiendo un indicador de confianza basado en la proporción de dicho acuerdo (número de votos para la clase mayoritaria/número de votos posibles), lo que permitió establecer un umbral de confianza superior a 80% (0,8) para cada predicción. Cada uno de los seis clasificadores, además del basado en la votación de los otros modelos, fue evaluado utilizando 30% del corpus de entrenamiento destinado a prueba, para comparar la clasificación manual de esa muestra con las predicciones arrojadas por los modelos.

Modelado profundo

Tras el modelado basado en aprendizaje superficial, se desarrolló una segunda estrategia para clasificar los textos basada en modelado profundo, usando *embeddings* como forma de representación del texto, y aprendizaje profundo, en concreto, redes neuronales recurrentes (RNN). De manera específica, se hizo uso del entorno de TensorFlow (v2) y Keras para crear un modelo secuencial con cuatro capas:

- La primera capa de entrada convierte cada palabra en *embeddings*, vectores densos que representan el valor categórico de cualquier palabra dada. Los *embeddings* se entrenaron utilizando las 10.000 palabras más comunes de la base creada, más 1000 de fuera de ese vocabulario. Por lo tanto, la matriz de *embeddings* incluyó una fila por cada una de estas 11.000 palabras y una columna para cada una de las seis dimensiones de *embeddings* (este híper-parámetro se ajustó varias veces y obtuvo el mejor rendimiento con tamaño = 6).

- La segunda y tercera capas incluidas son capas ocultas de unidad recurrente cerrada (GRU), y con 128 neuronas cada una. Los GRU son versiones simplificadas de las LSTM tradicionales, celdas de memoria a corto plazo duraderas usadas para crear redes neuronales recurrentes que permiten hacer predicciones en relación con secuencias de datos. Aunque ambas funcionan correctamente para clasificar texto (convergiendo rápidamente y detectando dependencias a largo plazo), se decidió aplicar GRU en lugar de LSTM porque la versión simplificada tiene un rendimiento similar, y al ser más simple ofrece una ejecución más rápida.

- La última capa de salida es la que permite la detección, y se trata de una capa densa con solo una neurona que usa la activación sigmoidea para predecir la probabilidad de que un mensaje contenga odio por cada una de las razones presentes en los corpus de entrenamiento.

Para compilar el modelo profundo, se usó *standard loss* con *crossentropy* binario y *adam optimizer*. Por último, se ajustó al corpus de entrenamiento para cinco épocas y se hizo uso de la parte del corpus destinado a prueba para la validación en treinta pasos. Ya que las redes neuronales requieren mucha capacidad de cómputo y existía la necesidad de escalar los procesos de lo local a lo distribuido, toda la recopilación de ejemplos, la clasificación manual y la generación de los modelos se ejecutaron de forma remota y paralelizada utilizando los servicios de cómputo del Centro de Supercomputación de Castilla y León.

RESULTADOS

Antes de revisar el rendimiento de los modelos generados, conviene analizar los resultados de la clasificación manual llevada a cabo para generar los corpus de entrenamiento.

Lo primero a señalar es que el porcentaje de tuits finalmente desechados es elevado, un 57,7% (N=13787) de la muestra, sumando los que no tenían acuerdo inter-codificadores y los descartados en la clasificación. También se observa que el porcentaje de tuits de odio motivado por la ideología política y validados con total acuerdo es reducido, pese a tratarse de mensajes previamente filtrados. Específicamente, y respondiendo a la PI1, se validó un 16,2% de tuits de odio (N=3879), frente a un 26,4% de mensajes de no odio (N=6334). Estas cifras evidencian, en primer lugar, que los diccionarios de filtros lingüísticos, por muy completos y complejos que sean, así como las técnicas de detección basadas en expresiones y palabras clave, no sirven como método eficaz para identificar los mensajes de odio en línea, algo que ya se presuponía. Sin embargo, sirvieron para acotar y optimizar el proceso, ya que sin esos filtros lingüísticos el trabajo para encontrar ejemplos de odio por ideología política en el histórico general de Twitter hubiese sido arduo. En segundo lugar, de estos datos podría concluirse que la cantidad de odio (al menos el explícito y el motivado por la ideología) que se propaga a través de Twitter no es tan importante como cabría esperarse, aunque quizás sí muy ruidosa y efectista. En la tabla 1, se puede consultar una muestra de tuits de ejemplo de cada una de las categorías resultantes.

Por otro lado, para evaluar los modelos predictivos generados se utilizaron tres de las métricas de evaluación más usadas en aprendizaje automático supervisado: *accuracy*, la media armónica *F-score* –que ofrece una métrica balanceada y calculada a partir de *precision* y *recall*– y *AUC-ROC* –que muestra el rendimiento de los modelos de clasificación en todos los umbrales de clasificación–. La tabla 2 muestra que todos los valores arrojados por estas métricas fueron aceptables, en la mayor parte de los casos por encima de 0,70. Comparando el rendimiento de cada uno de los algoritmos los valores de *accuracy* y *AUC-ROC* fueron considerablemente más altos en el modelo generado con redes neuronales recurrentes, lo que confirma la ventaja comparativa del aprendizaje profundo aplicado a clasificar textos. Así, respondiendo a la PI2, se puede concluir que, centrando la atención específicamente en el modelado superficial, el algoritmo de clasificación tradicional con mejor rendimiento en este caso es la regresión logística, seguido por el detector basado en los votos de los modelos superficiales y por Naïve Bayes Multinomial. Sin embargo, en términos generales, el modelo profundo ofrece un rendimiento mayor que los modelos generados con algoritmos superficiales (ver figura 5).

Tuits de *odio* con acuerdo
Anda a trabajar de una vez en tu vida, sucio comunista de mierda vende humos
Eres un cutre, puto facha asqueroso
La maldita izquierda y sus malditos delincuentes independentistas
Leña y más leña... ya esta bien de aguantar a #Guarros parasitos. #Asco #Izmierda
Mandemos a esa escoria socialista al gulag, hay que aplastarlos a todos nido de rojos de mierda
Muerte a los separatas y a los podemitas terroristas y narcocomunistas
Izquierdosos de mierda ojalá se mueran todos
Es que os metía a todos en una cámara de gas hijos de la gran puta
este tb es un hp. Un fascista de mierda el Rivera
INCREÍBLE como la perversa basura comunista pudre todo! Hay que acabar con ellos
Tuits de *no odio* con acuerdo
La izquierda en lo suyo como ya es costumbre
Las pancartas son siempre de los mismos
Si algunos ultras son de extrema izquierda, por qué nunca lo decís
okdiario Para ellos, un español es extrema derecha o no es nada
Estos comportamientos en Alemania estan castigado con cárcel! Vergüenza
Aquí un constitucionalista apoyando a los de la bandera del grajo, ya no se esconden
No tan solo la extrema izquierda gana, la extrema derecha esta al acecho...
el_pais Uy madre mía la que están liando las derechas ay Vox uy el trifachito
A3Noticias Puto montaje, que mediocridad y poca seriedad, buscan en sí populismo!
Este es el nivel de respeto de Vox, es decir, ninguno
Tuits sin acuerdo (mensajes ambiguos que generaron discrepancia)
Y esto es lo que pasa cuando se queman los contenendores, que la basura se acumula
Ojalá sigan cargándose el país estos comunistas que son lo mejor que nos ha pasado
Gran gestión de los amigos zurdos
Esos son los cómplices civiles de los crímenes de la derecha fascista
Vaya iagen: fachas de la estelada vs fachas de la bandera franquista. Qué absurdo y qué espanto
No se porque todavía me sorprendo con la habilidad de la izquierda de volver todo caos
Vaya, me encanta cuando el podemita le ronea a Marhuender
Antifascistas apoyando a un movimiento nacionalista, insolidàrio y xenófobo. No sé si serán antifascistas o solo idiotas
Claro que sí, que los islamoterroristas son niño de pecho al lado de la extrema izquierda liberal progres Globalistaa
Pues no se ya qué va a hacer Naranjito...buscar un espacio en la extrema izquierda

Tabla 1. Tuits de ejemplo de cada una de las categorías resultantes tras codificar

Fuente: Elaboración propia.

Aprendizaje superficial	*Accuracy*	*F-Score*	*AUC-ROC*
Naïve Bayes Original	,66	,73	,63
Naïve Bayes Multinomial	**,68**	**,78**	**,60**
Naïve Bayes Bernoulli	,62	**,76**	,50
Regresión logística	**,70**	**,78**	**,65**
Regresión lineal con gradiente descendiente estocástico	,67	,75	,63
Máquinas de vectores de soporte	,67	,75	**,64**
Modelo basado en los votos de los anteriores	**,70**	,75	**,64**
Aprendizaje profundo	***Accuracy***	***F-Score***	***AUC-ROC***
Redes neuronales recurrentes	**,81**	**,77**	**,89**

Tabla 2. Métricas de evaluación de los modelos generados con cada uno de los algoritmos

Fuente: Elaboración propia.

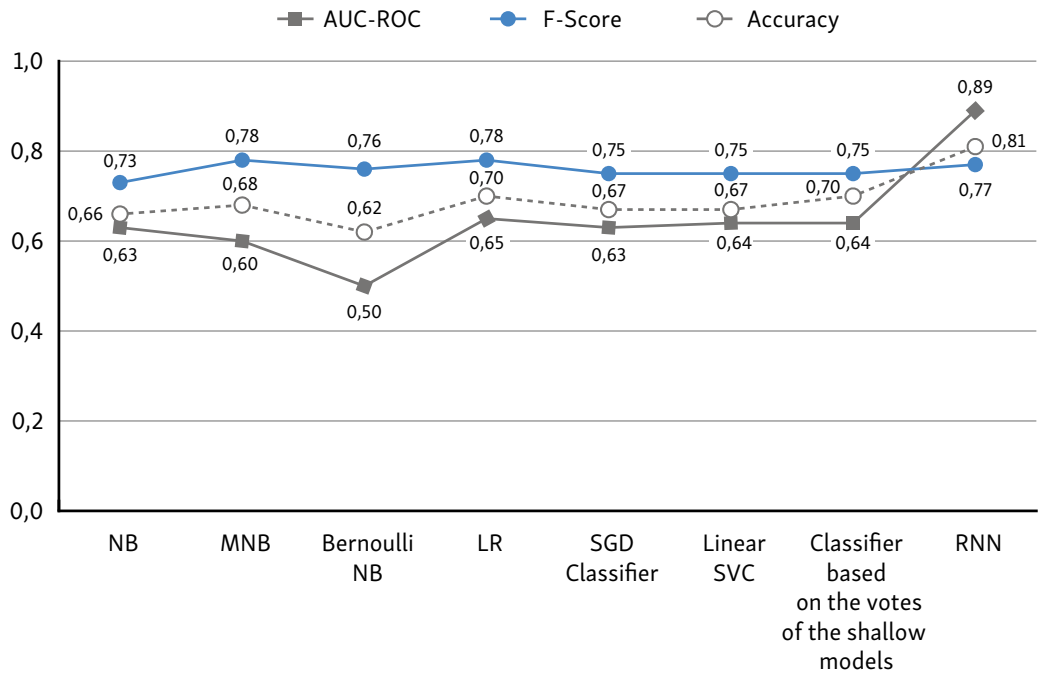

Figura 5. Métricas de evaluación de los modelos generados con cada uno de los algoritmos

Fuente: Elaboración propia.

CONCLUSIONES Y DISCUSIÓN

En este artículo se presenta el primer prototipo de detección automática de discurso de odio en Twitter en español motivado específicamente por razones de ideología política, modelado a partir de un corpus de entrenamiento generado manualmente ad-hoc, y haciendo uso, además, de aprendizaje profundo, una

innovación respecto de prototipos anteriores. Las principales técnicas usadas para desarrollar este prototipo han sido el procesamiento del lenguaje natural, para analizar y procesar datos no estructurados, y la clasificación de textos con aprendizaje automático supervisado, para detectar el odio por ideología política. La estrategia computacional desarrollada para el detector final implica descargar los mensajes desde la API *streaming* de Twitter y su procesamiento directo y masivo en el Centro de Supercomputación de Castilla y León, Scayle, donde se aplican los modelos predictivos que han sido entrenados y validados, para más tarde generar bases de datos con los mensajes finalmente clasificados con fiabilidad, en los grupos de odio y de no odio, para su observación por el usuario final.

Este trabajo confirma que es posible entrenar modelos predictivos que permitan detectar el discurso de odio en Twitter por un tipo de discriminación específico, como la ideología política, lo que además permite acotar y precisar mejor el entrenamiento de los modelos, resultando en un rendimiento sólido y una fiabilidad y precisión más que aceptables. Además, se ha creado una base de datos específica para entrenar los modelos predictivos, lo que permite mejorar la fiabilidad del detector aplicado a este contexto concreto, superando los posibles problemas de validez interna de anteriores prototipos. Cabe señalar que, aunque el porcentaje final de mensajes de odio y de no odio con acuerdo en el corpus de entrenamiento pueda parecer reducido, lo más importante en este proceso es contar con ejemplos de calidad, por encima de la cantidad. Esto se debe a que, aunque las métricas de evaluación pudieran ser aceptables, si los ejemplos no son completamente fiables la validez interna del prototipo se podría ver dañada, contaminada con falsos positivos o negativos. Por ello el foco se centró en generar un corpus de entrenamiento fiable y validado, ya que, además, la cantidad puede ser fácilmente ampliada con nuevos ejemplos clasificados.

En suma, se ha resuelto que, de los seis algoritmos de aprendizaje automático empleados en el modelado superficial, el que mejor rendimiento ofrece es la regresión logística, seguido por Naïve Bayes Multinomial. No obstante, en términos generales, se verificó que el aprendizaje profundo funciona considerablemente mejor que los algoritmos de clasificación convencionales para detectar este tipo de discurso de odio en Twitter, pues el modelo entrenado con redes neuronales presentó mejores métricas de evaluación.

Más allá de las cuestiones de índole técnica y metodológica, el estudio ha permitido observar una presencia notable de discurso de odio –16,2% de la muestra total y 38% de los tuits clasificados con certeza– sobre una muestra previamente seleccionada con palabras clave. Esto permite contribuir a las discusiones teóricas, no solo sobre su definición y taxonomía (Miró Llinares, 2016), los límites a la libertad

de expresión (Moretón Toquero, 2012) y las implicaciones (Müller & Schwarz, 2020), sino sobre su cuantificación. Esta tarea, especialmente compleja por la volatilidad de este discurso (Arcila Calderón et al., 2020), puede beneficiarse de una herramienta validada y específica como esta, para que se pueda medir el mismo tipo de odio –por ideología política– en diferentes periodos, contribuyendo a medir su evolución.

Se puede afirmar que este trabajo presenta un aporte metodológico, con la estrategia de detección a gran escala, la generación del corpus de entrenamiento ad-hoc y los modelos desarrollados con técnicas de aprendizaje automático supervisado; un avance teórico, en el estudio de los delitos de odio y, específicamente, del discurso de odio en Twitter por razones de ideología política, y una aplicación práctica, ya que la tecnología aquí desarrollada podrá ser implementada en diversos ámbitos públicos y privados. Esto último es lo de mayor relevancia, por su potencial aplicación por parte de las redes sociales para localizar y reducir la presencia de odio, por instituciones públicas, privadas o del tercer sector, incluyendo medios de comunicación e incluso partidos políticos, que intenten promover espacios menos radicalizados y polarizados. Así, el prototipo podría resultar de utilidad también en proyectos que buscan precisamente combatir el discurso de odio o la polarización en redes sociales, como WONT-HATE[1] o TRI-POL[2].

LIMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

A pesar de estas aportaciones, este proyecto presenta varias limitaciones a afrontar en un futuro próximo. En primer lugar, los modelos de clasificación desarrollados cuentan con métricas de evaluación aceptables. Sin embargo, el prototipo requiere de una validación que pruebe su fiabilidad al ser implementado de manera práctica, con nuevos casos reales, y siendo comparado con una nueva clasificación manual validada. Esto requeriría recopilar una nueva muestra de mensajes en un contexto diferente, codificarla manualmente de igual forma que se hizo con el corpus (dos codificadores) y, sobre esa misma muestra, correr los modelos y comparar los resultados de cada clasificación, pudiendo extraer después coeficientes de acuerdo entre la clasificación manual y la automática.

En segundo lugar, el prototipo desarrollado solo puede detectar el discurso de odio por ideología política en los mensajes de Twitter y en español, lo que

1. Liderado desde la Universidad de Navarra, financiado por el programa H2020. https://cordis.europa.eu/project/id/795937/es

2. Liderado desde la Universidad Pompeu Fabra, financiado por el Ministerio de Ciencia e Innovación y la Fundación BBVA: https://www.upf.edu/web/tri-pol

ha permitido desarrollar unos modelos más fiables, pero solo aplicables en este contexto, por lo que sería recomendable entrenar y desarrollar modelos basados en esta misma estrategia para detectar el discurso de odio en Twitter por otras razones de discriminación, así como en otros idiomas y contextos. Estas son tareas en las que los autores del artículo trabajan.

Finalmente, el prototipo se limita a detectar el discurso de odio únicamente en Twitter, por lo que conviene ampliarse a más fuentes, incluyendo medios sociales como YouTube o Instagram, blogs de medios, partidos políticos y asociaciones, así como plataformas digitales de los medios de información. En este sentido, aunque se reconoce que Twitter no es representativa de la opinión pública (ninguna red social de manera aislada), sus contenidos tienden a impactar y a viralizarse, alcanzando a todo tipo de personas, con o sin cuenta en la red social.

REFERENCIAS

Alonso González, M. (2017). Predicción política y Twitter: Elecciones generales de España 2015 (Political prediction and Twitter: Spanish legislative elections 2015). *ZER: Revista de Estudios de Comunicación = Komunikazio Ikasketen Aldizkaria, 22*(43), 13-30. https://doi.org/10.1387/zer.16298

Amores, J. J., Arcila-Calderón, C. A., & Stanek, M. (2019). Visual frames of migrants and refugees in the main Western European media. *Economics & Sociology, 12*(3), 147-161. https://doi.org/10.14254/2071-789X.2019/12-3/10

Amores, J. J., Arcila-Calderón, C., & Blanco-Herrero, D. (2020). Evolution of negative visual frames of immigrants and refugees in the main media of Southern Europe. *Profesional de la Información, 29*(6). Retrieved from https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/80525

Anti-Defamation League. (2020). *Online Hate and Harassment. The American Experience 2020*. The ADL Center for Technology and Society. Retrieved from https://www.adl.org/media/14643/download

Anti-Defamation League. (2021). *Online Hate and Harassment. The American Experience 2021*. The ADL Center for Technology and Society. Retrieved from https://www.adl.org/media/16033/download

Arcila-Calderón, C., Blanco-Herrero, D., & Valdez-Apolo, M. B. (2020). Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español (Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish). *REIS: Revista Española de Investigaciones Sociológicas, 172*, 21-40. https://doi.org/10.5477/cis/reis.172.21

Arcila-Calderón, C., Ortega-Mohedano, F., Amores, J. J., & Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático (Supervised sentiment analysis of political messages in Spanish: Real-time classification of tweets based on machine learning). *Profesional de la Información, 26*(5), 973-982. https://doi.org/10.3145/epi.2017.sep.18

Arroyo, S. C. (2017). El concepto de delitos de odio y su comisión a través del discurso: especial referencia al conflicto con la libertad de expresión (The concept of hate crimes and their execution through speech: special reference to the conflict with freedom of speech). *Anuario de derecho penal y ciencias penales, 70*(1), 139-225. Retrieved from http://agora.edu.es/servlet/articulo?codigo=6930585

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759-760). https://doi.org/10.1145/3041021.3054223

Bane, K. C. (2019). Tweeting the agenda: How print and alternative web-only news organizations use Twitter as a source. *Journalism Practice*, *13*(2), 191-205. https://doi.org/10.1080/17512786.2017.1413587

Benesch, S. (2014). Defining and diminishing hate speech. In P. Grant (Ed.), *State of the World's Minorities and Indigenous Peoples* (pp. 18-25). Retrieved from https://minorityrights.org/publications/state-of-the-worlds-minorities-and-indigenous-peoples-2014-july-2014/

Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. *Journal of Communication, 47*(1), 4-19. https://doi.org/10.1111/j.1460-2466.1997.tb02690.x

Carmona, O. I. (2010). Internet 2.0: El territorio digital de los prosumidores (Web 2.0: the digital territory of prosumers). *Revista Estudios Culturales*, (5), 43-64. Retrieved from http://servicio.bc.uc.edu.ve/multidisciplinarias/estudios_culturales/

Chetty, N. & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior, 40*, 108-118. https://doi.org/10.1016/j.avb.2018.05.003

Council of Europe. (1997). *Recommendation No. R (97) 20 of the Committee of Ministers to member states on "hate speech"*. Council of Europe, Committee of Ministers. Retrieved from https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680505d5b

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*. In *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14955

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). *Hate lingo: A target-based linguistic analysis of hate speech in social media*. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15041

European Commission against Racism and Intolerance. (2016). *ECRI General Policy Recommendation N.° 15 on Combating Hate Speech*. Council of Europe. Retrieved from https://book.coe.int/en/human-rights-and-democracy/7180-pdf-ecri-general-policy-recommendations-no-15-on-combating-hate-speech.html

Ferreira, C. (2019). Vox como representante de la derecha radical en España: un estudio sobre su ideología (Vox as representative of the radical right in Spain: A study of its ideology). *Revista Española de Ciencia Política*, (51), 73-98. https://doi.org/10.21308/recp.51.03

Jubany, O. & Roiha, M. (2018). *Las palabras son armas. Discurso de odio en la red* (Words are weapons. Hate speech online). Barcelona, Spain: Edicions Universitat Barcelona.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Paris, France: Unesco Publishing.

García-Ortega, C. & Zugasti-Azagra, R. (2018). Gestión de la campaña de las elecciones generales de 2016 en las cuentas de Twitter de los candidatos: entre la autorreferencialidad y la hibridación mediática (The management of the candidates' Twitter accounts in the Spanish 2016 general elections: Between self-referentiality and media hybridization). *Profesional de la Información*, *27*(6), 1215-1224. https://doi.org/10.3145/epi.2018.nov.05

Isasi, A. C. & Juanatey, A. G. (2017). *El discurso del odio en las redes sociales: Un estado de la cuestión* (Hate speech on social media: A state of the art). Barcelona, Spain: Ajuntament de Barcelona Progress Report. Retrieved from https://ajuntament.barcelona.cat/bcnvsodi/wp-content/uploads/2015/03/Informe_discurso-del-odio_ES.pdf

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559. https://doi.org/10.1108/IntR-06-2012-0114

Krippendorff, K. (2010). *On communicating: Otherness, meaning, and information*. Routledge.

Leader Maynard, J. & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70-95. https://doi.org/10.5038/1911-9933.9.3.1317

López-García., G. (2016). 'New' vs 'old' leaderships: the campaign of Spanish general elections 2015 on Twitter. *Communication & Society*, *29*(3), 149-168. https://doi.org/10.15581/003.29.3.149-168

López-Meri, A. (2015). Twitter como fuente informativa de sucesos imprevistos: el seguimiento de hashtags en el caso #ArdeValencia (Twitter as an Information Source of Unexpected Events: Following Hashtags in the Case #ArdeValencia). *Disertaciones: Anuario electrónico de estudios en Comunicación Social*, *8*(1), 27-51. https://doi.org/10.12804/disertaciones.01.2015.02

Malmasi, S. & Zampieri, M. (2017). *Detecting hate speech in social media*. arXiv preprint:1712.06427. Retrieved from https://arxiv.org/abs/1712.06427

Marín Dueñas, P. P. & Díaz Guerra, A. (2016). Uso de Twitter por los partidos y candidatos políticos en las elecciones autonómicas de Madrid 2015 (Use of Twitter by political parties and candidates in the 2015 Madrid regional elections). *Ámbitos: Revista Internacional de Comunicación,* (32), 1-15. Retrieved from https://revistascientificas.us.es/index.php/Ambitos/article/view/10436

Ministerio del Interior de España (Ed.). (2020*). Informe de Evolución de los Delitos de Odio en España* (Report on the Evolution of Hate Crimes in Spain). Retrieved from http://www.interior.gob.es/documents/642012/3479677/Informe+sobre+la+evolución+de+delitos+de+odio+en+España%2C%20año+2019/344089ef-15e6-4a7b-8925-f2b64c117a0a

Ministerio de Empleo, Migraciones y Seguridad Social. (2018). *Acuerdo de cooperación institucional con el Consejo General del Poder Judicial y la Fiscalía General del Estado, para luchar contra el racismo, la xenofobia, la LGBTIfobia y otras formas de Intolerancia* (Institutional cooperation agreement with the General Council of the Judiciary and the State Attorney General's Office, to fight against racism, xenophobia, LGBTIphobia and other forms of Intolerance). Retrieved from http://www.inclusion.gob.es/oberaxe/ficheros/ejes/cooperacion/Acuerdo_insterinsticuional_original.pdf

Miró Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet (Taxonomy of violent communication and the discourse of hate on the internet). *IDP. Revista de Internet, Derecho y Política*, (22), 82-107. Retrieved from https://www.raco.cat/index.php/IDP/article/view/n22-miro/408486

Mondal, M., Silva, L. A., & Benevenuto, F. (2017, July). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media* (pp. 85-94). https://doi.org/10.1145/3078714.3078723

Moretón Toquero, M. A. (2012). El «ciberodio», la nueva cara del mensaje de odio: entre la cibercriminalidad y la libertad de expresión (Cyberhate, the new face of the hate message: between cybercrime and freedom of expresión). *Revista Jurídica de Castilla y León, 27*, 1-18.

Movimiento contra la Intolerancia. (2019). *Informe Raxen: Racismo, Xenofobia, Antisemitismo, Islamofobia, Neofascismo y otras manifestaciones de intolerancia a través de los hechos. Especial 2019. Por un Pacto de Estado contra la Xenofobia y la Intolerancia* (Raxen Report: Racism, Xenophobia, Anti-Semitism, Islamophobia, Neo-fascism and other manifestations of intolerance through facts. Special 2019. For a State Pact against Xenophobia and Intolerance). Movimiento contra la Intolerancia. Retrieved from https://www.inclusion.gob.es/oberaxe/ficheros/documentos/InformeRaxen.pdf

Müller, K. & Schwarz, C. (2020). Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, jvaa045. https://doi.org/10.1093/jeea/jvaa045

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019). *Reuters Institute Digital News Report 2019*. Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf

Organization for Security and Cooperation in Europe. (2020). *OSCE - ODIHR. Hate Crime Reporting*. Retrieved from https://hatecrime.osce.org/

Pereira Kohatsu, J. C. (2017). *Construcción de modelos de clasificación automática para discursos de odio* (Building automatic classification models for hate speech) (Master's thesis). Retrieved from https://repositorio.uam.es/handle/10486/680053

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors, 19*(21), 4654. https://doi.org/10.3390/s19214654

Rodríguez, R. & Ureña, D. (2011). Diez razones para el uso de Twitter como herramienta en la comunicación política y electoral (Ten reasons to use Twitter as a tool for politicaland electoral communication). *Comunicación y pluralismo*, (10), 89-116. Retrieved from https://summa.upsa.es/viewer.vm?id=30573&view=main&lang=es

Said-Hung, E. M., Prati, R. C., & Cancino-Borbón, A. (2017). La orientación ideológica de los mensajes publicados en Twitter durante el 24M en España (The Ideological Orientation of Messages Posted on Twitter during the 24M in Spain). *Palabra Clave*, *20*(1), 213-238. https://doi.org/10.5294/pacla.2017.20.1.10

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences, 10*, 1. https://doi.org/10.1186/s13673-019-0205-6

Tamarit Sumalla, J. M. (2018). Los delitos de odio en las redes sociales (Hate crimes on social networks). *IDP: Revista de Internet, Derecho y Política*, *27*, 17-29. Retrieved from https://www.raco.cat/index.php/IDP/article/view/n27-tamarit

Valdez-Apolo, M. B., Arcila-Calderón, C., & Amores, J. J. (2019). El discurso del odio hacia migrantes y refugiados a través del tono y los marcos de los mensajes en Twitter (Hate speech against migrants and refugees through the tone and frames of Twitter messages). *Revista de la Asociación Española de Investigación de la Comunicación*, *6*(12). https://doi.org/10.24137/raeic.6.12.2

AGRADECIMIENTOS/FINANCIAMIENTO

Este trabajo fue desarrollado en el marco del proyecto *STOP-HATE. Desarrollo y Evaluación de un detector del discurso de odio en línea en español* [PC-TCUE18- 20_016], prueba de concepto competitiva liderada por el Dr. Carlos Arcila Calderón, y financiada por el Fondo Europeo de Desarrollo Regional y la Junta de Castilla y León a través del PLAN T-CUE de la Fundación General de la Universidad de Salamanca (2018-2020). Los autores agradecen especialmente su participación e implicación en el proyecto al Dr. Félix Ortega Mohedano y a todos los estudiantes de la Universidad de Salamanca que colaboraron en las tareas de codificación, sin las que hubiera sido imposible desarrollar este trabajo.

SOBRE LOS AUTORES

JAVIER JIMÉNEZ AMORES, es investigador miembro del Observatorio de los Contenidos Audiovisuales. Graduado en Comunicación Audiovisual y máster en Investigación en Comunicación Audiovisual por la Universidad de Salamanca, actualmente desarrolla su tesis doctoral en la misma Universidad, con el apoyo financiero de la Junta de Castilla y León y el Fondo Social Europeo. Su investigación se centra en el análisis de medios y redes sociales, la comunicación social, el discurso de odio y los métodos computacionales en ciencias sociales.

https://orcid.org/0000-0001-7856-5392

DAVID BLANCO-HERRERO, es doctorando en la Universidad de Salamanca, donde desarrolla su tesis con una beca FPU. Es graduado en Periodismo (Universidad a Distancia de Madrid) y Administración de Empresas (Universidad de León) y máster en Comunicación Audiovisual (Universidad de Salamanca). Es miembro del Observatorio de los Contenidos Audiovisuales y sus líneas de investigación son la ética periodística, la desinformación y el discurso de odio. Es asistente editorial en el Anuario Electrónico de Estudios en Comunicación Social "Disertaciones".

https://orcid.org/0000-0002-7414-2998

PATRICIA SÁNCHEZ-HOLGADO, es investigadora en la Universidad de Salamanca y miembro del Observatorio de los Contenidos Audiovisuales. Licenciada en Publicidad y Relaciones Públicas (Universidad Complutense de Madrid). Es profesora asociada en la Facultad de Lenguas y Educación de la Universidad Nebrija de Madrid y en la Facultad de Comunicación de la Universidad Pontificia de Salamanca. Experta en Big Data (Universidad Pontificia de Salamanca) y máster en Estudios de la Ciencia, la Tecnología y la Innovación (Universidad de Oviedo).

http://orcid.org/0000-0002-6253-7087

MAXIMILIANO FRÍAS VÁZQUEZ, es doctorando en la Universidad de Salamanca e investigador miembro del Observatorio de los Contenidos Audiovisuales. Licenciado en Ciencias de la Comunicación por la Universidad La Salle, México, y Máster en Investigación en Comunicación Audiovisual por la Universidad de Salamanca, investiga en las líneas de estudio de la migración y el discurso de odio, el análisis de redes sociales y el *big data*.

https://orcid.org/0000-0001-9750-6136