**UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE**

Programa de Doctorado en Criminología

TESIS DOCTORAL

# MICRO-ESPACIOS DE ODIO EN TWITTER

## Análisis de las características ambientales del ciberlugar para la detección y prevención de la comunicación violenta y el discurso de odio

Tesis doctoral presentada por

Zoraida Esteve Bañón

Dirigida por el Profesor Dr. D. Fernando Miró Llinares

Elche, diciembre de 2020

De acuerdo con la normativa de estudios de doctorado de la Universidad Miguel Hernández de Elche, se presenta como indicio de calidad de esta tesis doctoral el artículo con referencia "Esteve Bañón, Z., Moneva, A. y Miró Llinares, F (2019). Can metadata be used to measure the anonymity of Twitter user? Results of a Confirmatory Factor Analysis. *International E-Journal of Criminal Science, 13*, 1–16" (véase ANEXO II).

UNIVERSITAS
Miguel Hernández

D. Fernando Miró Llinares, Coordinador del Programa de Doctorado en Criminología de la Universidad Miguel Hernández de Elche, conforme a la normativa de Doctorado de la citada Universidad, presto conformidad y autorizo necesarias para que el trabajo de investigación presentado por Zoraida Esteve Bañón bajo el título "*Micro-Espacios de odio en Twitter. Análisis de las características ambientales del ciberlugar para la detección y prevención del discurso de odio*", pueda ser defendido como tesis doctoral con el fin de optar al grado de Doctor.

Atentamente,

FERNANDO|MIRO|LLINARES

Firmado digitalmente por FERNANDO|MIRO|LLINARES
Fecha: 2021.01.27 09:47:34 +01'00'

Fdo.: Prof. Dr. Fernando Miró Llinares

D. Fernando Miró Llinares, Doctor en Derecho y Catedrático de Derecho penal de la Universidad Miguel Hernández de Elche,

INFORMO

Que el trabajo de investigación presentado por Zoraida Esteve Bañón bajo el título *"Micro-Espacios de odio en Twitter. Análisis de las características ambientales del ciberlugar para la detección y prevención del discurso de odio"*, se encuentra en disposición de ser defendido como tesis doctoral con el fin de optar al grado de Doctor.

Atentamente,

FERNANDO|MIRO|LLINARES

Firmado digitalmente por FERNANDO|MIRO|LLINARES
Fecha: 2021.01.27 09:48:06 +01'00'

Fdo.: Prof. Dr. Fernando Miró Llinares

# INFORMACIÓN SOBRE LA FINANCIACIÓN

Esta tesis ha sido financiada por el Instituto Nacional de Ciberseguridad (INCIBE) en el marco de las “Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad” (ref. INCIBEI-2015-27349).

# AGRADECIMIENTOS

A mi maestro, Fernando Miró. Porque a su lado he crecido, profesional y personalmente, lo que nunca imaginé. Gracias por hacerme partícipe, desde el principio, de un gran proyecto como es Crímina. Y por haberme dado una de las cosas más importantes que tengo en mi vida. Estaré eternamente agradecida.

A las personas con las que comencé este camino. En especial a Francisco Bernabeu y José Eugenio Medina, por transmitirme su capacidad y tesón en el trabajo, así como su pasión por la Criminología. Gracias por enseñarme tanto. A Natalia García, María del Mar Ruíz y Elena Fernández, por su incesante apoyo, por las horas y horas de trabajo juntas y por compartir multitud de momentos que nunca olvidaré.

A todos mis compañeros que, con su esfuerzo y dedicación, continúan haciendo de Crímina un gran centro de investigación. De entre ellos, Fco. Javier Castro, Ana B. Gómez, Laura González, Flavia Roteda, María Dolores Gómez y Nahikari Sánchez. En especial, a Miriam Esteve, porque sin ella y su gran pericia en el ámbito informático, esta investigación no hubiera sido la misma. Pero, sobre todo, a Asier Moneva, por su tiempo y conocimientos, su compañerismo, su paciencia y sus palabras de ánimo constantes. Millones de gracias.

Al profesor Juanjo Medina, por la estancia de investigación en la Universidad de Manchester, en la que pasé horas y horas en sus preciosas bibliotecas.

Al profesor Roger Enriquez, por las dos estancias de investigación en la Universidad de Texas en San Antonio (EE.UU). Gracias por todas las conversaciones y sabios consejos. A Richard Hartley y Miguel Bedolla, por sus clases de historia. A mis amigos Gerardo y Carlos Rodríguez que estuvieron pendientes de mí en todo momento. A Henry, Mary y Maddie Meadde, por su hospitalidad, su cariño y por hacerme sentir en casa

incluso estando tan lejos de la mía. Fue una experiencia maravillosa que cambió mi vida por completo. Siempre seréis mi familia americana.

A mi familia y amigos. Gracias por entender los periodos de ausencia y apoyarme en este tedioso proceso durante tanto tiempo.

A mi hermano y, en especial, a mi sobrino David por su absoluta comprensión. Lamento no haberte dedicado todo el tiempo que me hubiese gustado. Te prometo que, a partir de ahora, te lo recompensaré.

A mis padres, Joaquín y Zoraida, por todo. Por su apoyo, esfuerzo y amor incondicional. Gracias por entender cuál era mi pasión y confiar en mí ciegamente.

A Álvaro, mi marido. Sin ti a mi lado no lo hubiera conseguido, sobre todo después de lo que hemos pasado durante este largo proceso. Gracias por poner tu pierna más fuerte cuando he cojeado y cubrir mi espalda con la tuya en todas las batallas.

"I'm with you till the end of the line"

(Captain America: The Winter Soldier)

# ÍNDICE

# RESUMEN

El fenómeno de la comunicación violenta y el discurso de odio en Internet posee unas características determinadas que lo convierte en potencialmente dañino para la sociedad. En este sentido, las redes sociales y—concretamente—Twitter se caracteriza por ser un entorno de gran comunicación e interacción entre sus usuarios, lo que facilita que los mensajes relacionados con la comunicación violenta y el discurso de odio alcancen a un mayor número de personas o colectivos. Este hecho, sumado a que se pueden mantener en el tiempo, potencia el daño a las víctimas. Además, este tipo de conductas llevadas a cabo en el ciberespacio puede desencadenar desórdenes sociales en el espacio físico. Por lo tanto, es fundamental desarrollar métodos para identificar y prevenir este fenómeno. A pesar de que las investigaciones relacionadas con la detección de la comunicación violenta y el discurso de odio en Internet han aumentado en los últimos años, se han centrado únicamente en el análisis semántico del contenido, excluyendo características ambientales que pueden ser relevantes para su estudio. En este sentido, la criminología ambiental ha demostrado ser de utilidad en el estudio y prevención del delito tanto en el espacio físico como en el ciberespacio, analizando las características del evento y de los ciberlugares para intervenir sobre ellos. Así, esta tesis doctoral pretende aportar un enfoque distinto al tradicional en el análisis de la comunicación violenta y el discurso de odio. De este modo, en la primera parte se analiza el fenómeno de manera exhaustiva y se muestran las herramientas de detección que se están empleando en la actualidad. En la segunda parte, se profundiza en los enfoques criminológicos del evento delictivo para la detección de la comunicación violenta y el discurso de odio en Internet centrando su análisis en los ciberlugares de Twitter. En la última parte, se lleva a cabo un estudio de las características ambientales de los mensajes publicados en la red social Twitter, con una muestra de tweets

recogidos después de tres atentados terroristas: El atentado a la revista satírica Charlie Hebdo en París en 2015, el atentado en el aeropuerto y metro de Bruselas en 2016 y el atentado en la ciudad de Londres en 2017. Esto ha permitido, por un lado, examinar las características y estimar la prevalencia real del fenómeno y, por el otro, identificar las diferencias entre las variables ambientales (metadatos) de los tweets neutrales y los relacionados con la comunicación violenta y el discurso de odio. Por último, se han empleado técnicas de *machine learning* para crear un modelo predictivo, basado en metadatos, capaz de identificar las diferentes características de la comunicación neutral que permite distinguirla de la comunicación violenta. El resultado obtenido facilita y reduce las tareas de análisis que realizan las autoridades públicas y los proveedores de servicios para identificar y prevenir este fenómeno en Twitter.

*Palabras clave: comunicación violenta y discurso de odio; prevención; detección; criminología ambiental; ciberlugares; metadatos; Twitter.*

# ABSTRACT

The phenomenon of violent communication and hate speech online presents certain characteristics that make it potentially harmful to society. In this sense, social media and specifically Twitter constitute an environment of great communication and interaction between their users, which facilitates messages related to violent communication and hate speech to reach a greater number of people or groups. This fact, added to the fact that they can be maintained over time, enhances the damage to victims. Furthermore, this type of behaviour carried out in cyberspace may trigger social disorders in physical space. It is therefore essential to develop methods to identify and prevent this phenomenon. Although research related to the detection of violent communication and hate speech online has expanded in recent years, it has focused mainly on the semantic analysis of content, excluding environmental characteristics that may be relevant to its study. In this sense, environmental criminology has proved to be useful in the study and prevention of crime both in physical space and in cyberspace, by analysing the characteristics of the event and the cyber places to intervene on them. Thus, this doctoral thesis aims to provide a different approach to the traditional one in the analysis of violent communication and hate speech online. In this way, the first part exhaustively analyses the phenomenon and shows the detection tools that are currently being used. In the second part, the criminological approaches to the crime event for the detection of violent communication and hate speech online are examined in depth, focusing on Twitter cyber places. In the last part, a study on the environmental characteristics of the messages published on the social media Twitter is carried out, using a sample of tweets collected after three terrorist attacks: The attack on the satirical magazine Charlie Hebdo in Paris in 2015, the attack on the airport and metro in Brussels in 2016 and the attack in the city of London in 2017. This allowed, on the one

hand, to examine the characteristics and estimate the real prevalence of the phenomenon and, on the other, to identify the differences between the environmental variables (metadata) of neutral tweets and those related to violent communication and hate speech online. Finally, machine learning techniques have been used to create a predictive model, based on metadata, capable of identifying the different characteristics of neutral communication that allow them to be distinguished from those of violent communication and hate speech online. The result obtained facilitates and reduces the analysis tasks carried out by public authorities and service providers to identify and prevent this phenomenon on Twitter.

*Keywords: violent communication and hate speech; prevention; detection; environmental criminology; cyber places; metadata; Twitter*

# MARCO TEÓRICO

# CRIMINOLOGÍA AMBIENTAL Y DETECCIÓN DE LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO EN INTERNET

# PARTE I

# LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO EN INTERNET Y SU PREVENCIÓN

# Capítulo 1

# LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO EN INTERNET

"Don't say anything online that you wouldn't want plastered on a billboard with your face (or logo) on it."

(Erin Bury, 2009)

Desde que las modernas tecnologías de la información y la comunicación irrumpieran en nuestras vidas, la difusión de ideas y sentimientos ha sufrido una profunda transformación. Hoy, los mensajes pueden alcanzar a cientos de miles de potenciales receptores, trasladando su contenido sin importar límites geográficos, políticos e incluso aquellos otros impuestos por el idioma. En consecuencia, junto a la comunicación cotidiana, inocua desde un punto de vista de los derechos de las personas, es posible propagar emociones, deseos o propuestas con capacidad para generar daño moral o violencia física. Sin embargo, no han sido las tecnologías en general o Internet en particular, los creadores de estas formas de comunicación, aunque sí la han modificado (Miró-Llinares, 2016).

En efecto, el ser humano, como ser social, ha mostrado a lo largo de la historia su necesidad de comunicarse y, en ese proceso, las palabras han sido causa de ofensas que han provocado violencia, enfrentamientos e incluso guerras. En ese sentido, la comunicación violenta y el discurso de odio se han limitado a concretos espacios físicos mientras no ha sido posible la comunicación a distancia. La aparición de la escritura primero y más tarde, la imprenta, con su capacidad para ampliar el alcance de los

mensajes a través de libros o folletos ha permitido, por ejemplo, la difusión de ideas de grupos extremistas o xenófobos. Así, se han ido sucediendo situaciones objeto de reproche cuya causa ha sido la difusión del odio en cualquiera de sus formas, encontrando el primer y máximo exponente en nuestro país, en los casos de las barcelonesas librerías Kalki o Europa, convertidas en el epicentro de la difusión del nacismo, razón por la que en el año 2016 fueron clausuradas (Miró-Llinares, 2016). Sin embargo, este estado de las cosas ha cambiado radicalmente tras la aparición y generalización de Internet. El nuevo espacio para la comunicación (Miró-Llinares, 2011), ahora permite la difusión de la comunicación violenta y el discurso de odio, no en una tienda física con cuya clausura se limita considerablemente la propagación de estas ideas, sino en un entorno sin ubicación ni fronteras, el Ciberespacio, al que acceden millones de usuarios a través de páginas web, blogs (Timofeeva, 2003) o de cualquiera de las formas de publicación de contenidos. Pero es que, además, tras el desarrollo de la web 2.0 y la consecuente creación y popularización de las redes sociales, las oportunidades de comunicación y acceso a los mensajes se han visto incrementadas notablemente. Internet se ha convertido, así, en un instrumento global para compartir opiniones y pensamientos entre individuos, lo que unido a la implementación de las herramientas de inteligencia artificial ha modificado nuestros patrones de consumo de contenidos y también, ha multiplicado los riesgos y experiencias negativas relacionadas con el cibercrimen en general y la difusión de la comunicación violenta y el discurso de odio en particular.

## 1. La Comunicación Violenta y el Discurso de Odio

No existe una definición universalmente aceptada sobre el discurso de odio (Brown, 2017; Howard, 2017; Martins et al., 2018; Matsui, 2016; Weber, 2009). La expresión,

traducción del término inglés *hate speech* acuñado a finales de 1980 en los EEUU (Brown, 2017), hace referencia a las opiniones que vierten juicios de valor negativos contra ciertos grupos y minorías sociales y sus miembros (Rey Martínez, 2015). Mari Matsuda lo utilizó por primera vez en 1989 en su artículo *Respuesta pública al discurso racista*, en el que trató de mostrar las dificultades y limitaciones que desde una perspectiva jurídica presentaba acotar este fenómeno. Tal fue la repercusión de su publicación que, tras la amplia aceptación por parte de otros juristas, terminó por ser adoptada por los legisladores y las autoridades judiciales, llegando incluso a extenderse a los medios de comunicación y a la población en general (Brown, 2017). También en España, algunos autores han llamado la atención sobre la necesidad de delimitar el significado de esta expresión. En este sentido, Fernando Rey (2015) ha criticado la terminología empleada, debido a que no representa de manera precisa lo que entendemos por "discurso de odio". De esta manera, aboga por la expresión "discurso discriminador", ya que el odio hace referencia a una cualidad interna de la persona que se puede convertir en una conducta discriminatoria, pero no siempre tiene que ser violenta. Por su parte, Fernando Miró (2016) va más allá y enmarca el discurso de odio dentro de una categoría más amplia denominada "comunicación violenta", que haría referencia a "cualquier acto comunicativo violento en Internet" (Miró- Llinares, 2016).

Esta falta de consenso tiene como consecuencia que, según sean las circunstancias y ámbitos en los que se aplique el término, se adopte una interpretación amplia o por el contrario más restrictiva. Por ejemplo, el Comité Europeo de Ministros del Consejo de Europa en su Resolución 20, de 30 de octubre de 1997, establece que el discurso de odio "abarca todas las formas de expresión que propaguen, inciten, promuevan o justifiquen el odio racial, la xenofobia, el antisemitismo u otras las formas de odio basadas en la intolerancia, incluida la intolerancia expresada por agresivo nacionalismo

y el etnocentrismo, la discriminación y la hostilidad contra las minorías, los inmigrantes y las personas de origen inmigrante". En este sentido, la incitación al odio abarca los comentarios dirigidos contra una persona o un grupo particular de personas (Weber, 2009).

La definición propuesta por el Comité Europeo de Ministros del Consejo de Europa en 1997 ha tenido una gran aceptación. Así, el Tribunal de Justicia de las Comunidades Europeas de Derechos Humanos (TEDH) la ha empleado en diferentes ocasiones. Esta versión restrictiva del concepto entiende que no es suficiente que las palabras simplemente ofendan, conmocionen o perturben a la víctima, sino que deben incitar al odio o a la violencia del resto de personas (Brown, 2017; Weber, 2009).

Como se ha indicado, no es la única propuesta, ya que otros organismos, también europeos, han adoptado su propia definición. Concretamente la Comisión Europea contra el Racismo y la Intolerancia (2016), describe el discurso de odio como "fomento, promoción o instigación, en cualquiera de sus formas, del odio, la humillación o en menosprecio de una persona o grupo de personas, así como el acoso, el insulto, los estereotipos negativos, la estigmatización o la amenaza con respecto a esa persona o grupos de personas y la justificación de todos los tipos de expresión anteriores, por motivos de raza, color, linaje, ascendencia, origen nacional o étnico, edad, discapacidad, idioma, religión o creencia, sexo, género, identidad de género, orientación sexual y otras características o condición personales". Como indica Howard (2017), esta última definición parece ser más amplia que las anteriores, ya que incluye acoso, insultos, estereotipos negativos y amenazas.

Una tercera opción es la propuesta por la Unión Europea en la Decisión Marco 2008/913/JAI del Consejo de 2008, en la que hace referencia a "toda conducta que incite públicamente a la violencia o al odio contra un grupo de personas o un miembro

de dicho grupo definido por referencia a la raza, el color, la religión, la ascendencia o el origen nacional o étnico (Consejo de la Unión Europea, 2008), excluyendo características tales como el sexo, la identidad de género y la orientación sexual".

Ante esta necesidad de concreción y la evidente dificultad de llevar a buen término la delimitación de la conducta por parte de las distintas instancias europeas, cuyo origen probablemente se sitúe en las habituales dificultades propias del dibujo de los límites de las conductas humanas, Tulkens (2013) añade que los diferentes planteamientos normativos de los países miembros hacen tremendamente difícil la adopción de una definición jurídicamente vinculante en Europa. No obstante, entiende que es posible identificar un denominador común en todos ellos, el discurso de odio tiene como objetivo comprometer la dignidad de aquellos a los que se dirige, tanto a sus ojos como a los de los demás miembros de la sociedad (Waldron, 2012), así como el deseo de atacar intencionadamente a una persona o a un grupo por motivos de raza, etnia, género, discapacidad, orientación sexual, religión o cualquier otro criterio prohibido. De este modo, de acuerdo con Tulkens (2013), debe existir voluntad en la ofensa.

La comunidad científica, por su parte, ha tratado con profusión el tópico. Por ejemplo, Nockleby (2000) define el discurso del odio como "cualquier comunicación que menosprecia una persona o un grupo sobre la base de algunas características tales como raza, color, etnicidad, género, orientación sexual, nacionalidad, religión u otra característica". Para Weber (2009), el concepto de discurso de odio podría referirse al: discurso de odio dirigido contra personas, o grupos de personas, pertenecientes a una raza concreta; el discurso de odio por motivos religiosos; y, por último, al concepto recogido por la redacción de la recomendación sobre el discurso de odio del Comité de Ministros de las Naciones Unidas.

Warner y Hirscherg (2012) entienden el discurso de odio como una forma específica de lenguaje ofensivo que hace uso de estereotipos para expresar una ideología del odio.

Por otro lado, Kaufman (2015) hace referencia a cuatro elementos básicos característicos del discurso de odio: (1) Situación de vulnerabilidad del sujeto, la cual debe estar relacionada con una discriminación histórica; (2) la agresión o humillación de símbolos que representen a un determinado grupo social; (3) la incitación para denigrar o humillar a colectivos en situación vulnerable; (4) la intención de excluir y humillar a las personas o colectivos concretos.

Para Brown (2017) el concepto de discurso de odio es complejo, y está compuesto por tres elementos básicos: (1) El habla u otra conducta expresiva; (2) los grupos o clases de personas identificadas por características protegidas; (3) y las emociones, sentimientos o actitudes de odio.

Teniendo en cuenta este análisis, el autor da gran importancia al concepto de odio, entendido como la emoción intensa e irracional de humillación, enemistad y desprecio hacia un individuo o grupo, contra el que se dirige por tener ciertas características reales o percibidas. Los términos como discriminación, homofobia, racismo o xenofobia, delimitan los distintos grupos en que pueden existir víctimas de estas acciones y que podrían concretarse en odio racial y étnico, odio por religión o nacionalidad y odio por género u orientación sexual (Esquivel, 2016).

En este sentido, basándose en las formulaciones de Brown (2017), Martins y sus colaboradores (2018) definen el discurso de odio, como cualquier expresión emocional que transmita opiniones o ideas subjetivas, dirigidas a un público externo, y con fines discriminatorios. Puede tomar muchas formas (escrita, visual, artística, etc.), y puede ser difundida a través de cualquier medio (prensa, radio o televisión), incluyendo Internet.

Para Assimakopoulos y sus colegas (2017), el término discurso de odio podría definirse como la expresión de odio hacia un individuo o grupo de individuos, sobre la base de características protegidas. Donde el término “características protegidas” implica la pertenencia a algún grupo social que podría, por sí solo, desencadenar la discriminación. Por el contrario, como indica Baider (2017), en un intento de definir discurso de odio de manera más amplia, se podría adoptar la perspectiva del Pacto Internacional de Derechos Civiles y Políticos, la cual no señala ninguna característica particular. Y, no obstante, propone que el discurso de odio equivale a una apología del odio discriminatorio que constituye incitación a la hostilidad, la discriminación o la violencia (Asamblea General de las Naciones Unidas, 1966). Lo que es más importante aquí, en todo caso, es la palabra "incitación", que ocupa un lugar central y hace que la intención de desencadenar acciones potenciales contra los miembros de grupos protegidos sea una condición previa para considerar un acto de incitación al odio. Asumiendo, por tanto, la existencia de un vínculo entre la incitación al odio y el delito de odio, de modo que la primera supuestamente conduce a la segunda.

Aunque se pretende llegar a un consenso terminológico, tanto por los organismos como por la comunidad científica, existen multitud de definiciones de la expresión discurso de odio o *hate speech*. Estas definiciones suelen formularse desde una perspectiva jurídica, indicando cuáles son las características que pueden convertir la conducta realizada en un hecho constitutivo de delito o, por el contrario, sea únicamente la consecuencia del derecho a la libertad de expresión. Algunas de ellas son más amplias, en otros casos, como hemos visto, son más restrictivas.

Sin embargo, no es objeto de esta investigación ensayar una nueva definición del discurso de odio. Tampoco realizar un análisis jurídico del fenómeno sino estimar su prevalencia para, posteriormente, elaborar un modelo predictivo que permita diferenciar

la comunicación neutral de la comunicación violenta y el discurso de odio a partir de sus variables ambientales. Así se podrá identificar y prevenir de manera más adecuada. Por tanto, de acuerdo con la taxonomía planteada por Miró-Llinares (2016), no nos centraremos únicamente en el estudio del discurso de odio, sino que abordaremos el fenómeno desde una perspectiva más amplia.

En este sentido, se entendería como comunicación violenta, cualquier forma de expresión o comunicación violenta en Internet, abarcando toda expresión que pueda concebirse como violenta, independientemente de que se realice por motivos discriminatorios (Miró-Llinares, 2016). Dentro de este macroconcepto se encontraría el discurso de odio propiamente dicho y la incitación a la violencia como forma específica de discurso de odio, caracterizada por incitar a la discriminación, la humillación al grupo y el deseo de un mal (Figura 1).


*Figura 1. Comunicación violenta y discurso de odio (Elaboración propia, 2020)*

## 2. Internet y La Comunicación Violenta y el Discurso de Odio

### *2.1. El orden de los factores…: La aparición de Internet y su impacto en la comunicación violenta y el discurso del odio*

La forma de comunicarnos ha evolucionado con la expansión y la popularización de Internet en general y las TIC en particular, afectando a todos los ámbitos de la vida. El uso de Internet ya forma parte de nuestra rutina diaria y sus herramientas nos permiten contactar con amigos y familiares distantes cientos e incluso miles de kilómetros, organizar viajes o realizar transacciones bancarias. El ciberespacio es una gran plataforma donde el usuario difunde y expresa libremente sus pensamientos y opiniones, con el potencial de transmitir cualquier publicación a millones de personas en un corto periodo de tiempo (Silva et al., 2016). La Red permite que los usuarios estén continuamente informados sobre cualquier suceso o tema de interés. En definitiva, Internet proporciona oportunidades de comunicación únicas con un coste casi nulo para los usuarios.

No obstante, este nuevo ámbito de comunicación también plantea inconvenientes. Entre otros, permite a aquellos que propugnan el odio, transmitir sus ideas universalmente (Hawdon et al., 2017). En este sentido, la expansión de la información en Internet ha permitido que determinados grupos extremistas puedan difundir su mensaje de manera incontrolada (Perry & Olsson, 2009). Debido a esta generalización de la tecnología ha habido un notable aumento de los grupos de odio *online* y las actividades relacionadas con el ciberodio en Internet (Perry & Olsson, 2009).

Si, hasta hace poco tiempo, en el espacio físico nos encontrábamos con mensajes privados que se quedaban en la intimidad de los implicados, en la actualidad nos enfrentamos a un fenómeno creciente (Sood & Churchill, 2012), una comunicación

ofensiva en el ciberespacio que puede convertirse en un riesgo social, afectando a valores tan importantes como la igualdad, la intimidad, la dignidad o incluso el orden público (Bautista-Ortuño, 2017). Además, en el ciberespacio podemos encontrar multitud de vídeos que incitan a la violencia, así como páginas web con mensajes de odio por motivos de orientación sexual, nacionalidad, de género, racial, etc. (Bautista Ortuño, 2017).

Una de las primeras expresiones de discurso de odio en la World Wide Web es la registrada el 11 de enero de 1995 con la aparición del foro neonazi *Stormfront* (Koppel, 1998.) Su fundador, el extremista Don Black, antiguo miembro del Ku Klux Klan, consiguió que esta web sirviera como un mercadillo de odio en la Red, dando voz a multitud de manifestaciones de racismo y antisemitismo (Cohen-Almagor, 2011), hasta que en 2017 fue bloqueada. Desde entonces, Internet se ha utilizado para promover la incitación al odio contra personas por su color, religión, orientación sexual, etnia, etc. Debido a la facilidad de acceso y a la rápida evolución tecnológica que ha vivido la sociedad, se ha incrementado la creación de sitios donde los grupos extremistas incitan al odio, comparten ideología, propaganda, enlaces a sitios similares e incluso reclutan a nuevos adeptos, abogando por la violencia y las amenazas (Cohen-Almagor, 2011; Tiven, 2003). En este sentido, la preocupación por parte de muchos países del mundo por el incremento del uso de Internet como herramienta para transmitir contenidos de odio, racistas y xenófobos, ha aumentado enormemente (Blaya, 2019; Cohen-Almagor, 2011).

Definir el ciberodio, y concretamente el *online hate speech*, es complicado dadas las variaciones culturales y lingüísticas (Burnap & Williams, 2015b). La definición que han adoptado en varios estudios Burnap y Williams (2015) es la formulada por Greenawalt (1989), quien afirma que cualquier análisis de la ley con respecto al discurso de odio en

el espacio físico debe considerar la medida en que este lenguaje tiene valor expresivo. Se deben considerar cuatro criterios que pueden hacer que tales expresiones puedan llegar a ser criminales:

a. Que puedan provocar una respuesta violenta.

b. Que puedan herir profundamente a aquellos a quienes va dirigido el discurso.

c. Que dicho discurso ofenda a aquellos que lo escuchan

d. Que las calumnias y epítetos tengan un efecto degradante en las relaciones sociales dentro de cualquier comunidad.

Este fenómeno puede tomar muchas formas y dirigirse, tanto a una persona como a un grupo, por características tales como su religión, raza, etnia, género, orientación sexual, nacionalidad, origen o alguna otra característica definitoria del mismo (Banks, 2010; Blazak, 2009; Näsi, 2015; Perry, 2009), pudiendo difundirse tanto por grupos organizados como por usuarios que actúan de forma independiente (Hawdon et al., 2014).

Por otro lado, comparte algunas similitudes con otros delitos cometidos en Internet, como puede ser el caso del ciberacoso. Pero, en el *online hate speech,* los materiales que se publican (fotos, texto, vídeo, etc.) difieren del resto de ciberdelitos. En este caso, expresan odio o actitudes degradantes hacia un colectivo característico, ya que no suelen atacar a los individuos de forma aislada (Räsänen et al., 2016).

En el año 2018, en España, se recogieron, por parte de las Fuerzas y Cuerpos de Seguridad, 166 casos (10 más que en el 2017) relacionados con el discurso de odio o *hate speech* y relativos a la orientación sexual e identidad de género, el racismo y xenofobia o la ideología. Los hechos delictivos denunciados más repetitivos fueron las injurias, amenazas y discriminación, siendo Internet (45,2%) y las redes sociales

(25,9%) las vías preferidas para la comisión de los mismos. Aunque, en menor medida, también se emplearon otros medios de comunicación como la telefonía/comunicaciones (13,3%) y los medios de comunicación social (4,8%) (Ministerio del Interior, 2018).

Y aunque, desde la perspectiva criminológica, podamos afirmar que el discurso de odio *online* en algunos casos puede no ser necesariamente un delito, es innegable que perjudica a las personas ofendidas (Silva et al., 2016). En este sentido, podría pensarse que el abuso verbal, cara a cara, es más perjudicial psicológicamente que el abuso verbal *online* (estando ambos basados en las mismas características concretas). Y ello quizás porque el primero es más personal, y el agresor muestra su identidad real. Pero, si pensamos en el efecto que puede tener el abuso *online*, siendo más frecuente, más repetitivo y frente a un público más amplio, parece que el discurso de odio *online* tiene un efecto especialmente dañino. Porque al tener una multitud de testigos, tiende a producir daños inmediatos y a corto plazo, incluyendo hiperventilación, dolor de cabeza, mareos, aumento de la presión arterial, incremento de las pulsaciones, consumo de droga en incluso el suicidio (Saunders, 2011), así como mayor vergüenza y angustia a la víctima (Brown, 2018). El impacto del discurso de odio no es el mismo en todos los casos, porque dependerá de la persona involucrada, el contenido, la ubicación y las circunstancias. El discurso de odio puede dañar a las víctimas directa o indirectamente, y en la incitación directa al odio, las víctimas resultan inmediatamente heridas por el contenido del mensaje (Chetty & Alathur, 2018) siendo un menoscabo de naturaleza psicológica y emocional (Leets & Giles, 1997).

En consecuencia, el comportamiento hostil en el ciberespacio no sólo tiene consecuencias perjudiciales para las víctimas (Keipi et al., 2017; Näsi et al., 2015; Tynes, 2006; Ybarra et al., 2008), sino que también puede considerarse como una amenaza para la inclusión social y una motivación potencial para delitos de odio en el

espacio físico (Awan & Zempi, 2016; Douglas, 2007; Waldron, 2012). En este sentido, existen investigaciones en las que se ha estudiado el discurso de odio como elemento desencadenante de agresiones en el espacio físico, como fue el caso de las elecciones de 2017 en Kenia (Gagliardone, 2014).

El contenido de odio y xenófobo, ampliamente presente en Internet, preocupa a los usuarios (Kaakinen et al., 2018) y también a los organismos encargados de las políticas internacionales (Consejo de Europa, 2015; Comisión contra el racismo y la intolerancia, 2016; Gagliardone et al., 2015), ya que promueve el prejuicio y el odio que, con el tiempo, puede afectar a las raíces de la sociedad, creando una división entre los grupos sociales y, en última instancia, provocar profundas divisiones en la cohesión social (Pálmadóttir & Kalenikova, 2018).

Por lo tanto, la pronta identificación y detección de la comunicación violenta y más concretamente el discurso de odio en la Red y su rápida eliminación, se convierte en una prioridad para prevenir consecuencias muy dañinas, tanto en las víctimas como en la sociedad en general.

### *2.2. Características de la comunicación violenta y el discurso de odio en Internet*

Existe una gran cantidad de literatura que, no sólo documenta la variedad y el alcance de la comunicación violenta en Internet, sino que también analiza si es diferente a la comunicación violenta offline (Citron, 2014; Citron & Norton, 2011; Cohen-Almagor, 2017; Delgado & Stefancic, 2014; Perry & Olsson, 2009; Tsesis, 2001). Para Brown (2018), las características que diferencian la comunicación violenta *online* de la offline son la facilidad de acceso, el tamaño de la audiencia, el anonimato y la instantaneidad. Siendo esta última la más distintiva para el autor. La naturaleza

instantánea de la comunicación *online* fomenta formas de ciberodio que son más espontáneas y, por lo tanto, más desconsideradas (Brown, 2018).

Algunas de las peculiaridades de Internet que podrían marcar potencialmente la comunicación violenta y el discurso de odio *online* como diferente del *offline*, serían (Brown, 2018): anonimato, invisibilidad, comunidad, e instantaneidad.

Respecto al anonimato, una de las características de Internet como medio de comunicación es que los usuarios no están obligados a revelar aspectos de su identidad en el espacio físico, a menos que lo deseen. (Graham, 1999) afirma que el anonimato puede proporcionar oportunidades para expresarse con libertad, ya que los usuarios pueden decir lo que piensan sin miedo a que otras personas reaccionen o respondan desfavorablemente, por sus características personales. Pero también se da la situación contraria, esto es, Internet desinhibe a las personas para decir cosas que de otro modo no dirían (Suler, 2004), o incluso animar a que sean más provocadoras o inquietas de lo que serían en el espacio físico (Branscomb, 1995; Citron, 2014; Coffey & Woolworth, 2004; Cohen-Almagor, 2015; Poland, 2016).

Una segunda característica diferenciadora en los discursos de odio *online* es la distancia física que puede existir entre los ofensores y las víctimas y la invisibilidad entre ambos. Los usuarios que expresan su odio en Internet actúan sin observar las señales sociopsicológicas normales de empatía y censura que suelen mantener bajo control el comportamiento perjudicial o antisocial. Es decir, si no se puede ver el daño emocional causado por el discurso de odio *online,* es más probable que se minimice su importancia, ya que los ofensores no pueden ver las caras de sus víctimas y del resto de usuarios que, sin ser víctimas, pueden desaprobar dicha conducta.

Otra de las características es el deseo innato de la gente de comprometerse con otras personas de ideas afines, y fomentar el sentimiento de comunidad. Personas que, de otra

manera, no serían capaces de interactuar debido a la distancia geográfica, o porque simplemente desconocen la existencia del resto (Posner, 2001).

En el ciberespacio, el tiempo que transcurre entre tener un pensamiento y expresarlo a otra persona particular, o a un grupo de personas (aunque se encuentren a gran distancia), puede ser cuestión de segundos. Sin embargo, en el espacio físico la distribución de la difamación es más complicada, e incluso se puede llegar a meditar tanto, que deje de llevarse a cabo. En este aspecto, Internet fomenta la espontaneidad y las reacciones viscerales del discurso de odio.

### *2.3. Clasificación de la comunicación violenta y el discurso de odio en Internet*

Levin y McDevitt (1993) formularon por primera vez una tipología de las motivaciones de los delincuentes de crímenes de odio, basándose en entrevistas con policías, víctimas y varios delincuentes de crímenes de odio (Tabla 1). Los autores desarrollaron una tipología que identificó tres motivos principales: delincuentes que cometieron sus crímenes por la excitación o la emoción, delincuentes que se veían a sí mismos defendiendo su territorio y, finalmente, un pequeño grupo de delincuentes cuya misión consistía en librar al mundo de grupos que se consideraban malvados o inferiores (McDevitt & Bennett, 2002). En un estudio actualizado por McDevitt y Bennett (2002), se examinaron las características de los tres motivos originales (la emoción, la defensa y la misión) en relación con una nueva categoría: la motivación de represalia. Basándose en estas investigaciones, Jacks y Adler (2015) realizaron un estudio que les permitió desarrollar una tipología original de discurso de odio en Internet, dividiéndola en cuatro tipos distintos de usuarios.

*Tabla 1. Tipología de las características del delincuente del discurso de odio en Internet (Jacks & Adler, 2015)*

| Tipos | Características |
|---|---|
| Navegadores | Las personas que navega por Internet y ven material de odio en el ciberespacio, pero no interactúan con la comunidad *online*. |
| Comentaristas | Las personas que, además de ver contenido de odio en Internet, publicarán comentarios y participarán con el resto de usuarios. |
| Activistas | Usuarios de Internet que agregan abiertamente contenido de odio en la Red y es más probable que estén involucrados con grupos de odio organizados en el espacio físico. |
| Líderes | Usuarios de Internet más serios y comprometidos. Tienen más probabilidades de infringir claramente la ley por medio de la incitación. Utilizarán Internet para apoyar, organizar y promover su ideología extremista. |

Esta tipología sugiere que los que difunden contenido de odio en Internet pueden clasificarse basándose, tanto en sus motivaciones, como en sus comportamientos *online* (Jacks & Adler, 2015). Por otro lado, Sharma y sus colaboradores (2018) confeccionaron una clasificación de comunicación violenta basándose en el grado de intención de odio (Sternberg, 2003). Los autores definieron tres clases que muestran varias categorías de diferentes tipos de lenguaje ofensivo:

Clase I: En esta categoría estarían incluidos los mensajes que incitan a acciones violentas, más allá del mismo discurso. Debe ser público o estar dirigido a un grupo particular. Esta clase haría referencia al tradicional discurso de odio más restrictivo.

Clase II: En esta categoría estarían recogidas las "bromas" en el ciberespacio en las que, por ejemplo, se usa un lenguaje agresivo/provocador. Aquí, la característica violenta es menor que en la Clase I, ya que no incita a una respuesta violenta. Debe ir dirigida a un grupo o comunidad.

Clase III: Sería el menor grado de intención de odio mostrado en las categorías. Hace referencia a los comentarios de naturaleza ligeramente provocativa, en su mayoría de manera individual, no necesariamente dirigida a un grupo o comunidad. El contexto gira principalmente entorno al trolling con un tono irónico y sarcástico.

Asimismo, Silva et al., (2016) categorizaron manualmente los objetivos de los discursos de odio que analizaron en la red social Twitter y Whisper. Por el ejemplo, el término "gay" se categorizó como orientación sexual y "negro" como raza. Las categorías que se obtuvieron fueron: Raza, Comportamiento, Físico, Orientación sexual, Clase, Género, Etnicidad, Discapacidad, y la Religión. También añadieron una categoría "Otra" para cualquier objetivo de odio no clasificado (Tabla 2).

*Tabla 2. Objetivos de odio para cada categoría (Silva et al., 2016)*

| Categorías | Ejemplos objetivos de odio |
|---|---|
| Raza | Negro, raza negra, gente blanca. |
| Comportamiento | Gente insegura con el comportamiento, gente sensible. |
| Físico | Gente obesa, gente guapa. |
| Orientación sexual | Gente gay, gente heterosexual |
| Clase | Gente de guetos, gente rica. |
| Género | Gente embarazada, gente sexista. |
| Etnicidad | Pueblo chino, pueblo indio, pakistaní |
| Discapacidad | Personas con retraso, bipolares. |
| Religión | Personas religiosas, personas judías. |
| Otra | Borrachos, gente superficial |

Por su parte Awan (2014) realizó un análisis de 500 tweets de 100 usuarios diferentes de Twitter para examinar cómo se percibía a los musulmanes y de qué

manera eran atacados a través de Twitter, ofreciendo una tipología de las características de los delincuentes (Tabla 3).

*Tabla 3. Tipología de las características del delincuente del discurso de odio en Internet (Awan, 2014)*

| Tipos | Características |
|---|---|
| El Pesquero | Persona que ha pasado por las cuentas de Twitter de otras personas para dirigirse específicamente a personas con una conexión musulmana |
| El Aprendiz | Alguien que es bastante nuevo en Twitter pero que, sin embargo, ha empezado a dirigirse a las personas con la ayuda de abusadores *online* más experimentados |
| El Diseminador | Alguien que ha tuiteado y retransmitido mensajes, fotos y documentos de odio *online* que tienen como objetivo específico a los musulmanes |
| El Imitador | Una persona que está usando un perfil falso, una cuenta y unas imágenes para dirigirse a los individuos |
| El Accesorio | Una persona que se une a las conversaciones de otras personas a través de Twitter para dirigirse a las personas vulnerables |
| El Reactivo | Una persona que después de un incidente importante, o problemas de inmigración, iniciará una campaña *online* dirigida a ese grupo e individuo específico |
| El Impulsor | Alguien que cambia regularmente su cuenta de Twitter para seguir apuntando a alguien de un perfil diferente |
| El Profesional | Una persona que tiene un gran número de seguidores en Twitter e independientemente de las consecuencias tiene y lanzará una gran campaña de odio contra un individuo o grupo de personas por ser musulmanes |

Como hemos visto, en la literatura científica criminológica existen diferentes definiciones (e.g. Greenawalt, 1989; Matsuda et al., 1993) y clasificaciones del fenómeno (e.g. Awan, 2014; Jacks & Adler, 2015; McDevitt & Bennett, 2002; Sharma et al., 2018; Silva et al., 2016). Pero, como ya se ha indicado, el propósito fundamental de esta investigación no es analizar la etimología del discurso de odio, ni siquiera examinar las cuestiones normativas o jurídicas del término, sino estimar su prevalencia, así como identificar las características ambientales de la comunicación neutral y

distinguirlas de la comunicación violenta y el discurso de odio. Elementos que ayudarán a identificar este fenómeno en las redes sociales para poder eliminarlo lo antes posible.

Partiendo de esta base, el concepto de odio que emplearemos para nuestra investigación será, como ya indicamos, el de Miró-Llinares (2016). La elaboración de su taxonomía nos parece la más adecuada ya que profundiza en su análisis, pasando del estudio único del discurso de odio al de la comunicación violenta. Parte de la idea básica de que la comunicación violenta, y el discurso de odio, van más allá de lo que tradicionalmente se ha denominado *hate speech* en el ciberespacio, ya que ese discurso inicial discriminatorio, que suele expresarse como incitación a la violencia, es una parte, pero no es la única que existe en Internet.

Por lo tanto, la idea principal de esta taxonomía es la de conceptualizar las diversas formas existentes de comunicación violenta en Internet (entre las que se encuentra el discurso de odio). Y que, de esta forma, se pueda identificar y clasificar cada una de las categorías distintas del resto. Para ello, trabajó con una muestra de 255.674 tweets obtenidos momentos posteriores al atentado terrorista del semanario satírico francés, Charlie Hebdo en 2015 (Miró-Llinares, 2016).

Esta clasificación es diferente al resto ya que, como hemos indicado, el objetivo fundamental de estudio no es exclusivamente el discurso de odio o *hate speech* sino todo acto de comunicación violenta en la Red, y no únicamente el discriminatorio. Además, tiene un fundamento valorativo-jurídico, debido a que aporta criterios para evaluar las decisiones sobre la criminalización y valorar las políticas preventivas que se puedan elaborar al respecto (Miró-Llinares, 2016). En este sentido, la violencia es la base fundamental de este tipo de comunicación, incorporando en su esencia la violencia física incitada, anunciada, deseada o justificada. También la violencia moral, así como

la resultante de un daño no físico, o de una ofensa en el ámbito moral de los intereses personales individuales o colectivos (Miró-Llinares, 2016).

Por lo tanto, según indica Miró (Miró Llinares, 2016), existen otras formas de comunicación violenta diferentes al *hate speech*:

a. Cualquier forma de incitación (directa o indirecta), o amenaza concreta de causar un mal a través de la violencia física, que no sea debida a la discriminación o a la pertenencia de grupo con unas características concretas.

b. Cualquier daño, o menosprecio, al honor o a la dignidad de personas concretas.

c. Cualquier comportamiento que pueda estimarse ofensivo, o vejatorio, para la sociedad, aunque no vaya orientado a una persona concreta.

Teniendo como referencia este planteamiento, Miro-Llinares (2016) elabora una clasificación que se divide en dos grandes grupos:

a. En el primero se recoge todo discurso referido a la causación de un daño físico.

b. En el segundo, el discurso que ofende o causa un daño moral, tanto a individuos como a grupos sociales específicos o al resto de la sociedad.

Esta clasificación se configura sobre la base del término violencia, en la que se introduce una perspectiva valorativa (Figura 2). Lo que, además de clasificar los mensajes, permite valorar los discursos por sus consecuencias, indicadas en las cinco categorías (Miró-Llinares, 2016): α) incitación/amenaza directa a la violencia física; β) enaltecimiento a la violencia física; γ) ataques contra el honor o la dignidad; δ) incitación a la discriminación; ε) ofensas a la sensibilidad colectiva. Las cinco macrocategorías agrupan las distintas modalidades de comunicación violenta y de odio

que se pueden encontrar en la comunicación humana cotidiana (Castro & Bautista, 2019), si bien nuestro ámbito de estudio se centrará en la identificación y detección de la comunicación violenta y el discurso de odio en Internet para prevenirlo o reducirlo.

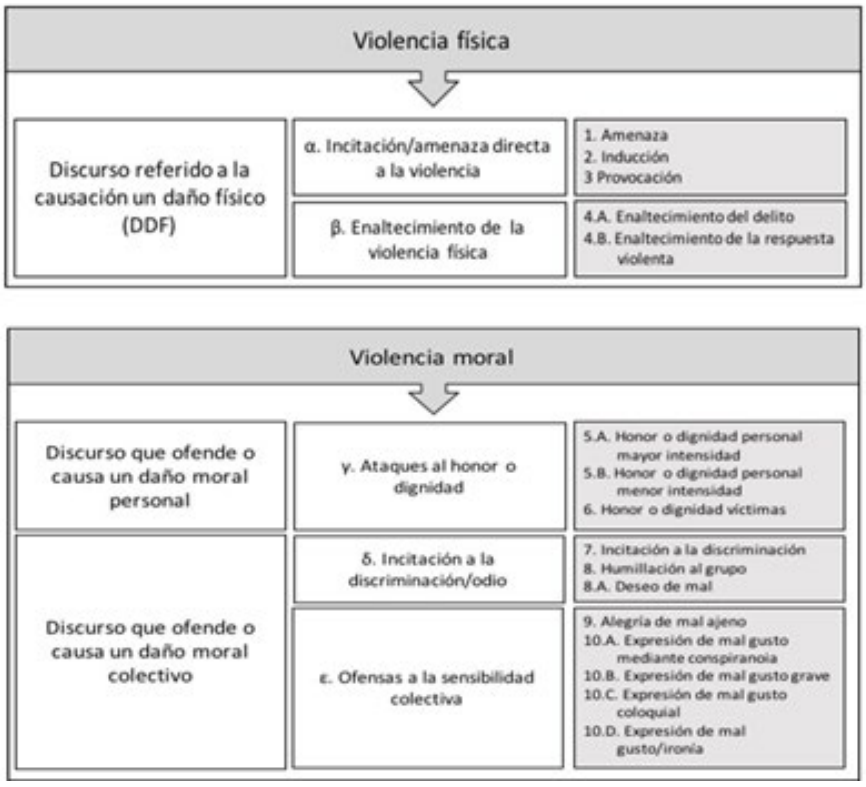


Figura 2. Taxonomía de la comunicación violenta y discurso de odio (Miró-Llinares, 2016)

## 3. La Evolución de la Digitalización

Aunque la comunicación violenta y el discurso de odio no es un fenómeno nuevo, sí ha ido evolucionando a la vez que lo ha hecho Internet. Debido a la modificación que ha sufrido en función de las fases evolutivas del ciberespacio, se han creado características propias que hacen necesario su análisis.

### *3.1. La primera fase*

La web 1.0, referida como la primera generación de la World Wide Web, se implementó en 1989 y duró hasta el 2005. Se definió como la red de conexión de información (Khanzode & Sarode, 2016). Los consumidores podían intercambiar información de manera universal, aumentando el acceso a la comunicación por parte de infinidad de usuarios. En este sentido, Internet se convirtió en un magnífico transmisor de información. Con anterioridad a esta revolución digital, los textos escritos que podían incluir pensamientos o discursos personales ya fueran de odio o no, se plasmaban en libros que se ponían a disposición de la sociedad a través de librerías o bibliotecas. Con la llegada de la web 1.0 la barrera de la distancia física desaparece y el mensaje puede llegar a cualquier parte del mundo (Pollock, 2006 citado en Miró-Llinares, 2016), como observaron los grupos radicales.

Aunque la Web 1.0 se caracterizó por tener contenidos únicamente de lectura disponibles en páginas web estáticas para cualquier usuario en cualquier momento, sin ningún tipo de interacción ni contribución en su contenido, fue suficiente para aquellos que querían difundir sus ideas violentas y de odio al resto de personas. En este caso, el administrador del sitio era el único responsable de la gestión del contenido.

### *3.2. La web 2.0 y 3.0*

Una de las características del ciberespacio, es el cambio permanente al que está expuesto. De este modo, la web 1.0 evolucionó a la web 2.0, dando más protagonismo a las personas, siendo más participativa y colaborativa en las actividades diarias de las mismas (Choundhury, 2014). Así, comienzan a surgir diferentes herramientas y maneras de utilizar Internet que fomentan la interacción entre los usuarios. De este modo, nacen las Redes Sociales (Castañeda et al., 2011). Con la nueva versión 3.0 las redes sociales,

siguen evolucionando, modernizando la gestión de los datos, fomentando el acceso a través del móvil y, en general, mejorando la satisfacción de los clientes.

De manera paulatina, las redes sociales se han instaurado en nuestras vidas cambiando la forma de interactuar con el resto de la sociedad, tanto a nivel personal como profesional. De hecho, en el año 2020, en todo el mundo, existen 3.805 millones de usuarios activos en redes sociales, un 83% más que en el año 2015 (Figura 3).

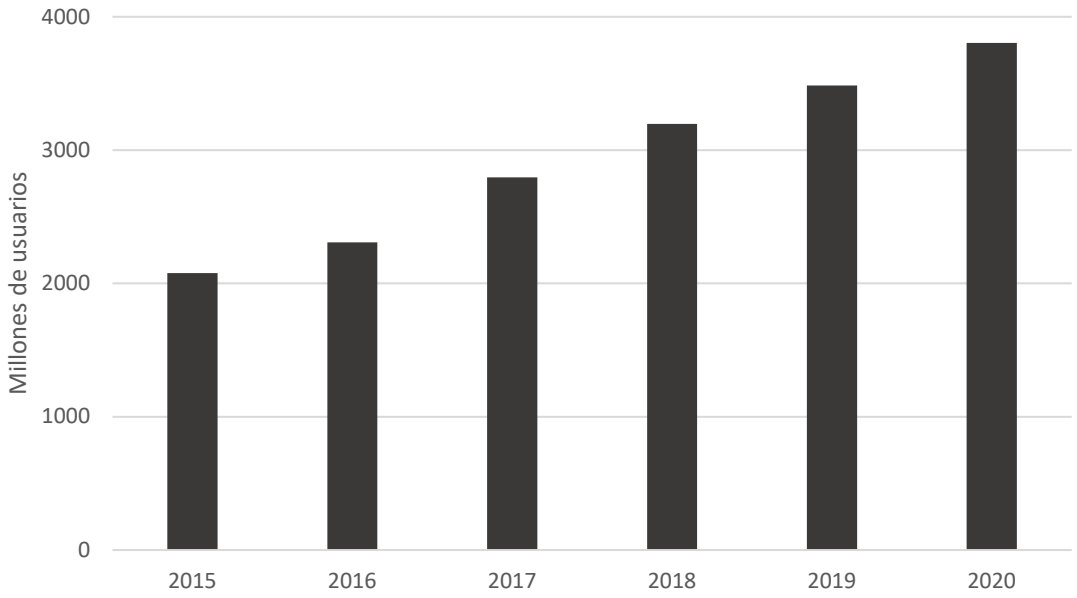


*Figura 3. Nº de usuarios de redes sociales, en el mundo, por año (en millones) con variación interanual (Elaboración propia a partir de datos de "we are social" y "hootsuite", 2020)*

La población en España en enero del 2020 era de 47.431.256 millones de personas (INE, 2020), de las cuales alrededor de 43 millones son usuarios activos de Internet, es decir, un 93%. Además, 29 millones son usuarios activos de las redes sociales, es decir un 62% de la población total. De estos usuarios, el 98% accede a través del móvil (we are social, 2020). La utilización de las redes sociales se ha extendido de manera muy dinámica, ya que permite al usuario mantener contacto con otras personas que están lejos geográficamente o que tienen intereses comunes (Antón-Prieto & Calderón, 2011) de una manera económica, rápida y, en algunos casos, divertida (Espinar-Ruiz &

González-Río, 2009). Además, ayudan en la divulgación de la información y el conocimiento compartido (Sánchez & Pinochet-Sánchez, 2017).

En general, parece que todos son ventajas, pero su uso también fomenta determinados comportamientos que pueden ser reprochados, como es el caso de los haters y su discurso de odio *online* (hate speech).

En definitiva, si lo analizamos, las redes sociales no aportan nada nuevo en la esencia de la comunicación de las personas, pero han conseguido una forma más avanzada de interaccionar entre ellas, evolucionando y adaptándose a las necesidades de la era digital (Merodio, 2010).

### 3.2.1 *¿Qué Entendemos por Redes Sociales?*

La definición a la que más se alude en la mayoría de investigaciones (Rohani & Hock, 2009; Allen, 2008; Flores-Cueto et al., 2009; Griffith & Liyanage, 2008; Kent, 2008; Lockyer & Patterson, 2008) es la enunciada por Boyd y Ellison (2007). Para las autoras, las redes sociales se definen como servicios basados en la web que permiten a los individuos:

- Construir un perfil público o semipúblico dentro de un sistema delimitado.
- Crear una lista de otros usuarios con los que comparten relaciones.
- Ver y recorrer su lista de relaciones y las realizadas por otros usuarios dentro del sistema.

El principal objetivo de los participantes de estas redes no es buscar gente nueva en su vida, sino comunicarse con personas que ya forman parte de ella, como pueden ser familiares o amigos (Boyd & Ellison, 2007). Para generar un perfil, normalmente se deben incluir datos como la edad, lugar de residencia o intereses personales. Y se anima

al usuario a subir una foto de perfil, dejando a la elección del mismo la discreción de su cuenta.

El trabajo de Boyd y Ellison ha sido la base para las investigaciones de otros autores. De este modo, Subrahmanyama et al. (2008), definen las redes sociales como una herramienta de comunicación *online* que permite a los usuarios la creación de perfiles públicos, o semipúblicos, visibles para ellos y para sus amigos que están registrados en esa misma red, así como su interacción entre ellos (Linda Castañeda et al., 2011).

En la misma línea de Boyd y Ellison, encontramos la definición de Benevenuto et al. (2009), quienes argumentan que las redes sociales sirven para que los usuarios se conecten entre sí, busquen y difundan información, y la compartan. Asimismo, la definición de Children International (2008) se basa en las redes sociales como lugares sociales creados para facilitar la interacción entre los usuarios a través de la colaboración, la comunicación y el intercambio de contenidos.

Por otro lado, Kaplan y Haenlein (2010), argumentan que las redes sociales son aplicaciones que facilitan al usuario la conectividad a través de la creación de perfiles personales, gracias a los que pueden ser vistos y ver a otros usuarios, y comunicarse a través de mensajes.

Teniendo en cuenta que la definición de Boy y Ellison (2007) surgió el año en el que Facebook se expandió a nivel mundial y en el que Myspace se convierte en el lugar más popular de Internet. Castañeda y Sánchez (2010) cuestionan que este tipo de redes hayan evolucionado y por ello sostienen que dicha definición necesita una actualización. De este modo, coinciden con las investigaciones de Beer (2008) y Fuchs (2009), donde indican que se debe avanzar en las definiciones propuestas y en las clasificaciones de las herramientas que han surgido, en relación con Internet, pensando que las redes sociales

forman parte de las herramientas que se generaron con la Web 2.0. La cual, como ya sabemos, ha evolucionado a la 3.0 (Castañeda & Sánchez, 2010).

Así, destacan la definición que se enuncia en el informe de INTECO (2009) en el que consideran las redes sociales como "servicios prestados a través de Internet que permiten a los usuarios generar un perfil, desde el que hacer públicos datos e información personal y que proporcionan herramientas que permiten interactuar con otros usuarios y localizarlos en función de las características publicadas en sus perfiles". Gracias a dos matices de esta definición, como la creación de un perfil y la interacción entre los participantes, podemos diferenciar las redes sociales de otros lugares Web (Castañeda et al., 2011).

Otra definición es la propuesta por De Haro (2010) , en la que entiende las redes sociales como los servicios que facilitan la construcción de un perfil propio y en la que se pueden crear conexiones con otros usuarios de la red, comunicarse entre ellos y donde el servicio no se especializa en algo específico sino en intereses generales.

Siguiendo esta tendencia, Castañeda y sus colaboradores (2011) afirman que una adecuada definición de las redes sociales sería "aquellas herramientas telemáticas de comunicación que tienen como base la Web, se organizan alrededor de perfiles personales o profesionales de los usuarios y tienen como objetivo conectar secuencialmente a los propietarios de dichos perfiles a través de categorías, grupos, etiquetados personales, etc., ligados a su propia persona o perfil profesional" (Castañeda & Gutiérrez, 2010) (p.18), siendo ese atributo el que les distingue del resto de herramientas.

Por otro lado, encontramos la definición de Prieto-Gutierrez (2011)entendiendo la red social como "una organización o estructura generada a través de las relaciones de diferentes actores (personas, instituciones, organizaciones, sociedades, etc.), debiendo

poseer o estar vinculadas a ciertas particularidades o rasgos comunes con el fin de poder interactuar entre sí".

En esta línea, también se sitúa Ponce (2012), que define las redes sociales como:

Estructuras sociales compuestas por un grupo de personas que comparten un interés común, relación o actividad a través de Internet, donde tienen lugar los encuentros sociales y se muestran las preferencias de consumo de información mediante la comunicación en tiempo real, aunque también puede darle la comunicación diferida en el tiempo, como en el caso de los foros. No sólo nos relacionamos y compartimos con los demás, sino que, además, exponemos abiertamente y en tiempo real nuestros gustos y tendencias, expresando la propia identidad. (p.6).

La autora defiende que las redes sociales facilitan a sus usuarios exponer sus intereses, fotografías, estados o vídeos, pero siempre sometidos a las condiciones de uso de dicha red. Es cierto que cada usuario puede configurar la privacidad de su cuenta en relación a la conexión de otros miembros, la visibilidad del perfil, alternativas de búsqueda de amigos, pero siempre dentro de las condiciones que están predeterminadas por la red social en concreto. En este sentido, las funciones de las redes sociales y sus utilidades son muy similares: creación de cuentas, personalización del perfil, envío de solicitudes de amistad con otros usuarios, etc. En definitiva, son los propios usuarios los que hacen que las redes sociales funcionen, cuando realizan el registro y aceptan la solicitud de amistad de otros usuarios para interactuar con ellos, compartiendo información y los contenidos generales de sí mismos.

Según Ponce (2012) las actividades sociales que se desarrollan en las redes sociales serían las siguientes:

- Compartir contenidos como fotografías, vídeos, textos, música, páginas web o noticias.
- Enviar mensajes privados a otros usuarios.
- Participar en juegos sociales que ofrece el servicio.
- Comentar el contenido compartido por otros usuarios.
- Crear grupos exclusivos para determinados contactos.
- Publicar eventos para anunciar acontecimientos a su red de amigos.
- Hablar en tiempo real con uno o más usuarios mediante chat o sistemas de conversión grupal.

Aunque las redes sociales han ido evolucionando desde que se originaron, estas características, junto con las opciones para configurar la privacidad de la cuenta, fomentan el uso y las relaciones de los usuarios en las mismas, estableciendo el funcionamiento general de sus servicios (Ponce, 2012).

### 3.2.2 *Historia y Evolución de las Redes Sociales*

El origen de las redes sociales es difuso al carecer de unidad de criterios sobre cuál fue la pionera del mercado. Si retrocedemos en el tiempo, encontramos que, en 1971, Ray Tomlinson envía el primer e-mail entre dos ordenadores, situados uno al lado de otro, en pleno progreso de ARPANET (Advanced Research Projects Agency Network). En 1978, Randy Suess y Ward Christensen fundaron Bulletin Board Bystems (BBS) cuyo objetivo principal era compartir noticias e información con sus amigos (Ponce, 2012). En 1994 y 1995 se lanzan GeoCities y The Glob respectivamente, ofreciendo un servicio que permitía a los usuarios configurar sus propios lugares web y personalizar sus experiencias en la Red, publicando el contenido y compartiendo intereses comunes.

En 1995 también se crea Classmates, por Randy Conrads, a modo de red social para interactuar con antiguos compañeros de la escuela. En 1997 se fundó Instant Messenger (AOL), el cual ofrecía un servicio de mensajería instantánea (Casado-Riera, 2017). Pero teniendo en cuenta la definición de redes sociales que hemos empleado, podríamos considerar que la primera red social que se lanzó al mercado en Estados Unidos fue "SixDegrees.com", en ese mismo año. Esta red social permitió a los usuarios crear perfiles, confeccionar una lista con amigos y, a partir de 1998, navegar por las listas de los amigos, promocionándose como una herramienta para ayudar a las personas a conectarse a Internet y enviar mensajes a otros usuarios. Aunque consiguió atraer a millones de usuarios, no consiguió convertirse en un negocio rentable y, en el año 2000 el servicio se cerró[1] (Boyd & Ellison, 2007).

En 1999 aparecieron LiveJournal, AsianAvenue y BlackPlanet. Herramientas que fomentaban la confluencia entre los perfiles de los usuarios y la relación con amigos, al igual que MiGente en el año 2000. Estas redes sociales permitieron a los usuarios crear perfiles personales, profesionales y de citas.

Por otro lado, en Corea se creó Cyworld en 1999. Pero no se convirtió en una red social (SNS) hasta el año 2001. Del mismo modo, la comunidad web sueca LunaStorm se remodeló como SNS en el año 2000, contenía listas de amigos, libros de visitas y páginas de diarios. Le siguieron otros lugares de redes sociales (SNS) como Ryze.com, lanzado en 2001 para ayudar a los usuarios a aprovechar sus redes de negocios. Los creadores de Ryze.com (2001), Tribe.net (2003), LinkedIn (2003) y Friendster (2002) estaban estrechamente relacionados tanto a nivel personal como profesional. Por ello, su filosofía era la de apoyarse entre ellos, sin competir. Sin embargo, Ryze no consiguió

[1] Según Boyd y Ellinson (2007), el 11 de julio de 2007 el creador de *SixDegrees*, A. Weinreich, afirmó en una comunicación personal que su red social estaba adelantada a su tiempo, porque la mayoría no contaba con amplias redes de amigos que estuvieran conectados

gran popularidad, Tribe.net creció para atraer a una base de usuarios apasionados por los mismos temas y LinkedIn se convirtió en un servicio de negocios muy potente.

A principios de 2004 apareció *Facebook*, pero únicamente para los estudiantes de la Universidad de Harvard, ya que, para unirse a esta red social, el usuario debería tener un correo electrónico de dicha universidad. A partir de septiembre de 2005, Facebook se amplió para incorporar a profesionales, estudiantes de secundaria y finalmente, a todo el mundo. Este mismo año, aparece Youtube. Y, en 2006, la red social de microblogging, Twitter. Junto a Bla Bla Car, Badoo, Waze y Tuenti. En el año 2008 se fundó Tumblr, otra red social de microblogging y al año siguiente nacieron Foursquare que permitía compartir la localización del usuario en las redes sociales y Runtastic, que analizaba la actividad deportiva del usuario y la interacción social con el resto de amigos. Google no podía quedarse atrás y en 2010 fundó su propia red social, Google Buzz. Y se crearon Instagram y Pinterest para compartir vídeos y fotografías. Al año siguiente, Google creó de nuevo otra red social, Google+. En el año 2012 se lanzaron Vine y Tinder. Y, en 2004, se creó Snapchat. En el año 2015 nació la aplicación de Twitter que permitía transmitir vídeos por streaming: Periscope. Y se creó Bebee, una red social española de búsqueda de empleo. En el año 2017 se lanzó Amazon Spark, dirigida a los miembros Prime de Amazon, cuyo objetivo era básicamente presumir de lo que se compra. Dependiendo del contenido que se genere en ella, puede salir recompensado.

Por último, se lanzaron varias redes sociales, entre ellas Farecast, en la que se pueden crear vídeos cortos de 15 a 60 segundos, usar opciones de videochats o transmitir en streaming; Tik-Tok (anteriormente Musical.ly), se basa en la subida de vídeos musicales y cortos divertidos que se comparten entre los usuarios y Lasso, creada por Facebook y copia de Tik-Tok, en la que se puede generar, editar y compartir vídeos cortos que tengan efectos especiales (Tabla 4).

*Tabla 4. Línea temporal de las redes sociales*

| Año | Redes Sociales |
|---|---|
| 1997 | SixDegrees.com |
| 1999 | LiveJournal; AsianAvenue; BlackPlanet |
| 2000 | LunaStorm (SNS relanzamiento); Mi gente; SixDegrees.com cierra |
| 2001 | Cyworld; Ryze.com |
| 2002 | Fotolog; Friendster; Skyblog |
| 2003 | Couchsurfing; LinkedIn; MySpace; Tribe.net; Open BC/Xing; Last.FM; Hi5; POF |
| 2004 | Orkut; Dogster; Flickr; Piczo; Mixi; Facebook (solo Harvard); Multiply; aSmallWorld; Dodgeball; Care2; Caster; Hyves; Viadeo |
| 2005 | Yahoo!360; YouTube; Xanga; Cyworld (China); Bebo (relanzamiento); Facebook (solo utilizado en el Instituto); AsianAvenue; BlackPlanet (relanzamiento) |
| 2006 | QQ (relanzamiento); Facebook (red corporativa); Windows Live Spaces; Cyworld (EE.UU); Twitter; MyChurch; Facebook (para todo el mundo); Badoo; Waze; Bla bla car |
| 2008 | Tumblr; Moterus |
| 2009 | Foursquare; Runtastic |
| 2010 | Google Buzz; Pinterest; Instagram |
| 2011 | Google + |
| 2012 | Vine; Tinder |
| 2013 | Telegram |
| 2014 | Musical.ly; Snapchat; WhatsApp |
| 2015 | Periscope; Bebee |
| 2017 | Amazon Spark |
| 2018 | Facecast; Hero Traveler; Tik-tok (anteriormente Musical.ly); Lasso (Copia de Tik-tok creada por Facebook) |

### 3.2.3 *Clasificación de las Redes Sociales*

Las características de las redes sociales pueden ser muy diversas en función de los intereses de los usuarios (ocio, viajes, relaciones sociales, trabajo, etc.). Aunque existen diferentes clasificaciones (por localización geográfica, sujeto principal, plataforma utilizada, etc.), nuestro análisis se va a centrar en la más generalizada ateniendo al público objetivo y a la temática.

#### 3.2.3.1 Redes sociales horizontales

Las redes sociales horizontales son aquellas que no tienen una temática concreta, abiertas a todo tipo de personas con diferentes intereses. De Alsola (2008) define este tipo de redes como "aquellas dirigidas a todo tipo de usuarios y sin temática definida".

- YouTube. La red social "YouTube" nació en febrero de 2005 con el objetivo de compartir vídeos entre los usuarios y, en 2006 la adquirió Google Inc. Actualmente se ha convertido en un fenómeno universal gracias a la cantidad de material audiovisual disponible que se comparte en todo el mundo. En España, en el año 2019, aumentó la frecuencia de su uso: el 59% de usuarios entra a diario a esta red social, mientras que en 2018 era del 53% (Iab, 2019). El éxito de esta plataforma digital reside en la facilidad de su uso para visualizar su contenido e incluso crear, editar y emitir material propio a través de un canal propio, en el que otros usuarios pueden realizar comentarios al respecto. Los usuarios asiduos a esta red social audiovisual que publican vídeos continuamente con el propósito de informar, entretener, expresar sus ideas, o contar sus experiencias referentes a cualquier temática, reciben el nombre de youtubers. Algunos de ellos incluso viven de su actividad en la red ya que, gracias a su gran capacidad para llegar al público, son reclamados por multitud de compañías que quieren darse a conocer o publicitarse entre la comunidad virtual.

- Facebook y Twitter. En el año 2004, Mark Zuckerberg creó una red social exclusiva para los estudiantes de la Universidad de Harvard. Y, aunque en septiembre de 2005 se amplió sólo para estudiantes de secundaria y profesionales, en 2006 se extendió al resto del mundo. El objetivo de su creador fue facilitar el contacto con la familia y los amigos, tener información de los sucesos en el mundo y compartir experiencias y comentarios entre los usuarios. Aunque con anterioridad a su creación ya existían algunas redes sociales, la aparición de Facebook cambió radicalmente este concepto (Lincoln & Robards, 2014), ya que -gracias a su buscador- te permite hacer un rastreo de amigos, o gente conocida, para interactuar con ellos a través de noticias, fotografías, comentarios, etc. Facebook también ha sido objeto de duras críticas entorno a su privacidad y transparencia, ya que juega con la delgada línea entre lo público y privado, con acciones que se pueden realizar en la propia red, como la publicación de contenido (comentarios, fotografías, etc.), sobre usuarios que no han dado los permisos correspondientes. Además, se ha cuestionado que esta red social sirva como repositorio de datos para la investigación (Zimmer, 2010), volviendo a la preocupación sobre la privacidad. En cualquier caso, sigue siendo una de las redes sociales con más usuarios en nuestro país (87%) (Iab, 2019). En el capítulo 4, nos centraremos en la red social de microblogging Twitter.
- Instagram. La red social Instagram se fundó en 2010 por Kevin Systrom y Mike Krieger. Su objetivo era subir fotos y vídeos de corta duración para compartirlos con los amigos o seguidores. Aunque se diseñó únicamente para iPhone, en 2012 se lanzó también para Android. Su atractivo reside en las numerosas herramientas para editar y retocar las fotos, y los vídeos, con filtros propios de esta aplicación. Pero, sobre todo, en la sencillez de su uso, ya que no se necesita ninguna cámara

especial, ni conocer una técnica especial de manejo. Estas publicaciones pueden ser privadas, pero la mayoría de los usuarios tienen cuentas públicas. Por lo tanto, es fácil encontrar y seguir a quien se desee. Al igual que los youtubers, existen los instagramers, que serían los usuarios de esta red social que comparten su vida a través del contenido visual que generan. Existen algunos instagramers con gran influencia en la comunidad digital. Por eso, las marcas y empresas más populares los utilizan para hacer publicidad de sus productos a través de sus seguidores.

En España, Instagram ha sido la red social que más ha crecido en el número de usuarios en el año 2019, con un porcentaje del 54% frente a un 49% en el año 2018 (Iab, 2019).

#### 3.2.3.2 Redes sociales verticales

Las redes sociales verticales son aquellas que se caracterizan por tener una temática común, facilitando la comunicación entre los usuarios que tiene los mismos intereses, y con una mayor especialización que las redes sociales horizontales (De Haro, 2010).

- Redes sociales verticales profesionales

Están dirigidas a generar relaciones profesionales entre los usuarios. Los ejemplos más representativos son: Viadeo (que ya no está operativa), Xing, Bebee y Linked In, que se ha constituido en la mayor red profesional del mundo con más de 645 millones de usuarios en más de 200 países y territorios (LinkedIn, 2020).

En España, la red social por excelencia en el ámbito profesional también es Linked In, con una frecuencia diaria de uso del 23% (Iab, 2019). Fue creada en el año 2002 por Reid Hoffman, y se lanzó en mayo del 2003 para facilitar a los usuarios la búsqueda de empleo y las relaciones profesionales. El usuario debe darse de alta en la red introduciendo su Curriculum Vitae (CV). De esta manera puede enviar sus referencias a

cualquier persona que esté en la misma red de contactos. Además, los contactos pueden realizar una recomendación sobre la opinión personal que tienen del usuario, aumentando la información del CV. Por lo tanto, se dirige al ámbito profesional, animando a los usuarios a que confeccionen su CV y realicen contactos que les puedan interesar, creando grupos que incluyen empleados de empresa, redes de estudiantes o grupos de interés (Sánchez & Pinochet-Sánchez, 2017). Es una buena herramienta para los usuarios que buscan trabajo y para los profesionales de recursos humanos que buscan perfiles profesionales muy concretos.

- Redes sociales verticales de citas

Del mismo modo que nuestra forma de comunicarnos ha evolucionado con las TICs, la manera de ligar se ha adaptado a las nuevas tecnologías. Así, buscar pareja, amistades o encuentros ocasionales a través de la Red se ha convertido en un método cada vez más popular en el ciberespacio. Existe una amplia oferta en redes de este tipo, ya sean de pago (Meetic, eDarling, etc.) o gratuitas (Badoo, Pof, Happn, Adopta a un tío, etc.).

Pero sin duda, una de las más utilizadas en España es Tinder con un 52 % de frecuencia de uso (Iab, 2019).

Tinder se lanzó en los EEUU en el año 2012 gracias a su creador Sean Rad, una persona introvertida con problemas para relacionarse (Jiménez-Acosta, 2017). La finalidad de esta red social era la de simular las situaciones que se pueden dar en el espacio físico que favorecen el flirteo. Su uso es sencillo, ya que en un minuto se puede configurar el perfil, aunque para registrarse se debe hacer con perfil de Facebook. El usuario debe redactar un texto breve sobre sí mismo y seleccionar un máximo de seis fotos. También debe delimitar el radio de distancia que le interesa, así como el sexo y el rango de edad de las personas que quiere conocer. La aplicación, con los datos

obtenidos, comienza a analizar el perfil localizando a usuarios que puedan ser compatibles y muestra sus perfiles.

Por lo tanto, únicamente se debe decidir si le interesa (deslizando el dedo a la derecha) o no (deslizando el dedo a la izquierda). Si las dos personas coinciden y se gustan, se les notifica (si no hay coincidencia, Tinder no informa cuando existe un rechazo) y comienzan a escribirse en un chat privado.

Desde su creación hasta la actualidad ha ido evolucionando en función de las necesidades que han ido detectando. Por otro lado, como todas las aplicaciones, ésta también tiene sus riesgos, como mantener relaciones sexuales con desconocidos. Por lo tanto, los usuarios deben tener las mismas precauciones que las empleadas en el espacio físico.

- Redes sociales verticales para viajar

Estas redes sociales están muy extendidas entre los usuarios de la Red que quieren servicios y/o información de sus viajes. Dentro de la misma temática, encontramos diferentes tipos.

- Bla, bla, car. Esta red social surgió en Francia en el año 2009, convirtiéndose en líder a nivel mundial en viajes con vehículos compartidos con más de 80 millones de usuarios en 22 países, incluida España (desde el año 2010) con 5 millones. La aplicación facilita el contacto entre personas que desean realizar un viaje en coche al mismo lugar y en las mismas fechas. De este modo, pueden compartir gastos de gasolina y peaje (Bla, bla,car, 2020).
- Waze. Waze es una aplicación de tráfico y navegación GPS en la que los usuarios comparten información sobre el estado del tráfico en tiempo real con el principal objetivo de evitar atascos. También permite previsualizar las

mejores rutas indicando la hora de llegada. Y, gracias a la acción colaborativa de los usuarios, informa de los servicios disponibles durante el viaje, si hay accidentes, obstáculos, obras, etc. (Waze, 2019). En España, en el año 2019, aumentó la frecuencia de su uso a diario (32%) respecto al año 2018 (28%) (Iab, 2019).

- Hero Traveler**.** Hero Traveler es una red social de reciente creación (2018) dirigida a todas las personas que han evolucionado al ritmo de las nuevas tecnologías, no recurren a las agencias de viajes para gestionar sus vacaciones (billetes, reservas, alojamiento, etc.) y se informan de los destinos a través de otras redes sociales (Instagram, Pinterest, etc.). En definitiva, facilita que cada viajero cuente la historia de su viaje para informar al resto de usuarios.

- Redes sociales verticales de ocio

Su objetivo es congregar a colectivos que desarrollen actividades de ocio, deporte, lectura, música, etc.

- MySpace**.** MySpace nació en 2003, pero tuvo su auge máximo entre los años 2005 y 2008, siendo la red con más visitas del mundo en esos años. Dejó de tener protagonismo, ya que los usuarios preferían otro tipo de redes sociales como Facebook. El formato evolucionó y se ha reconvertido en una red social especializada en música, donde los usuarios comparten su afición por ella y se dan a conocer artistas noveles.

- Runtastic**.** Esta red social se lanzó al mercado en el año 2009, siendo su objetivo principal registrar los datos del entrenamiento de los usuarios en actividades como running, ciclismo, senderismo, etc., empleando las

posibilidades del GPS para posteriormente analizarlos. Los usuarios pueden compartir sus logros con el resto, pueden fijarse objetivos propios o participar en los retos propuestos por otras personas, lo que ayuda a la motivación en grupo.

- Moterus**.** Moterus nació en 2008, creada por Francesc Pla e Isaac Feli, con el objetivo de unir al colectivo motero en una red social. Además de crear el perfil común a cualquier red social, subiendo fotos y vídeos, se caracteriza por tener actividades offline, como organizar y compartir rutas de moto y realizar quedadas para salir a la carretera en grupo. En general, es una red social que facilita la comunicación y la interacción entre los aficionados a las motos.
- Dogster**.** Esta red social se creó en el año 2004 para los amantes de los animales, y cuyas mascotas ocupan un lugar muy importante en su vida. Lo curioso de Dogster es la estructura de la web, donde los protagonistas no son los usuarios, sino las mascotas. Existen apartados para subir videos, fotografías, consejos y un apartado que ofrece animales en adopción.

### 3.2.4 *Los Hijos de las Redes Sociales*

El ciberespacio en general y las redes sociales en particular han creado nuevos fenómenos de culto para muchos usuarios, sobre todo para los más jóvenes.

#### 3.2.4.1 Influencers

Este término se emplea para definir a aquellos usuarios de las redes sociales que han conseguido una gran popularidad entre sus seguidores, ya sea por su imagen, credibilidad o estilo, creando tendencia en su modo de vida. De esta manera, las grandes marcas los utilizan como herramienta, o reclamo publicitario, gracias a la

capacidad que tienen de influir en las decisiones de sus seguidores (Gillin, 2007). Existen muchas clases de influencers y por diferentes motivos. Y, por eso, el marketing relacionado con este nuevo fenómeno se ha extendido muy rápidamente. De hecho, se han creado empresas y agencias que median entre las marcas y los influencers de manera estratégica, fomentando una nueva de forma de negociar, en la que los influencers cobran un gran protagonismo a la hora de publicitar determinadas marcas (Del Pino & Galán, 2010).

Así, además de los personajes famosos (actores, cantantes, deportistas, modelos, etc.) que, tradicionalmente, han sido determinantes en las campañas publicitarias, también nos encontramos con comunicadores de las redes sociales que, siendo anónimos, han conseguido que sus contenidos sean virales, alcanzando una gran difusión en el contenido multimedia en que publican. Keller y Berry (2003) argumentan que los influencers tienen múltiples intereses, son los primeros en obtener determinados productos y poseen la confianza de sus seguidores.

Dependiendo de la red social en la que se desarrolle el influencer, tendrá distinto nombre:

- YouTubers**.** Los YouTubers son creadores de contenido que dedican gran cantidad de tiempo a grabar, editar y subir los vídeos a YouTube (Holmbom, 2015) para compartirlos en la Red. La diferencia con el resto de usuarios es que ellos no lo hacen por hobbie, sino de manera profesional.
- Bloggers de moda**.** Son personas que se dedican al mundo de la moda, la ropa y otros aspectos de la industria. En los blogs que escriben y publican ofrecen consejos de compra, informan sobre las tendencias de moda, juzgan y dan su propia opinión sobre los productos. Según Corcoran y colaboradores (2006) , se pueden dividir en diferentes categorías:

- Insiders: Serían aquellos que trabajan o han trabajado en la industria de la moda.
- Outsiders: Tienen gran interés por la moda, pero su trabajo se desarrolla en otro ámbito.
- Aspirantes a Insiders: Serían aquellas personas que desean entrar en el negocio de la moda, ya sea como diseñadores, editores de revistas o columnistas.
    - Instagrammers. Los instagrammers son aquellos usuarios de la red social Instagram que tienes una gran influencia sobre sus seguidores. De este modo, las empresas de marketing los reclaman para publicar fotos y vídeos con sus productos y ampliar sus campañas publicitarias.

El funcionamiento de bloggers e instagrammers es muy parecido, ya que publican las fotografías o los vídeos con enlaces directos a las marcas que publicitan (ropa, complementos, etc.). La diferencia es que en Instagram el efecto del contenido de esa publicación todavía es más directo que en el blog, ya que en este último se indica toda la información de las marcas al final de la entrada. Por lo tanto, el lector tiene que hacer el esfuerzo de llegar hasta el final del texto si quiere conocer cuáles son las características e información de esos productos. En Instagram, a diferencia del anterior, existen "etiquetas" insertadas en el propio contenido audiovisual, las cuales permiten acceder directamente a la información relacionada a través de un enlace de hipertexto.

#### 3.2.4.2 Trolls y haters

El término troll hace referencia a los usuarios, de Internet en general y de las redes sociales en particular, que se dedican a provocar a otros. Schwartz (2008) definió a los trolls como personas normales que hacían locuras en Internet. Por su parte, Herring et al. (2002) se refirieron al trolling como un comportamiento que incluye atraer a otros a discusiones sin sentido y que consumen mucho tiempo. En la misma línea, Donath

(Donath, 1995) declara que el trolling es un juego sobre el engaño de identidad, aunque se juega sin el consentimiento de la mayoría de los jugadores y añade que puede ser costoso para los usuarios de la comunidad.

Algunos autores (Wilcox, 1998) argumentan que la palabra "troll", cuando se usa para referirse a personas que intentan provocar a otros, podría haberse originado en el ejército estadounidense en la década de los 60, al estar documentado el uso del vocablo trolling por parte de los pilotos de la Armada de los Estados Unidos para describir a la amenaza de los cazas rusos MIG en Vietnam. Otros autores (Crystal, 2006) argumentan que el término “troll” deriva de la tradición escandinava de los “trolls” como personajes horribles que merodeaban bajo los puentes.

Por otro lado, el término “trolling” resulta controvertido porque en sus inicios hacía referencia a la publicación de mensajes provocativos, y actualmente se ha extendido para referirse a la publicación de mensajes ofensivos (Bishop, 2013) y a un comportamiento destructivo, perturbando a los usuarios sin un propósito u objetivo aparente (Buckels et al., 2014). En 2011, se popularizó para describir el abuso en Internet tras una serie de trágicos casos de “R.I.P Trolling”, en los que se atacaron las páginas conmemorativas de personas fallecidas (Bishop, 2014).

De este modo, los trolls se divierten generando el caos en la Red. A pesar de la conciencia social existente relacionada con este fenómeno, la investigación empírica sobre el trolling es escasa, aunque en los últimos años se ha incrementado.

Schachaf y Hará (2010) realizaron entrevistas a trolls de Wikipedia e identificaron los factores que les motivaban para desarrollar su actividad. El aburrimiento fue el motivador más común por los entrevistados, ya que éstos sugirieron que simplemente desean divertirse mientras interactúan con otros usuarios. Además, entre sus motivaciones, incluyeron la búsqueda de atención, la venganza, el placer y el deseo de

causar daño a la comunidad. Del mismo modo, en los resultados de los análisis observaron que el comportamiento de los trolls se caracterizaba por realizar acciones repetitivas, intencionadas y dañinas que se llevaban a cabo de manera aislada y bajo identidades virtuales ocultas.

Por su parte, Hardaker (2010) llevó a cabo un análisis de contenido de determinadas publicaciones identificando cuatro características del trolling: la agresión, el engaño, la interrupción de la actividad y el éxito.

En general, el trolling consiste en la publicación de mensajes a través de una red que pretenden ser provocativos, ofensivos o amenazantes (Bishop, 2013). El uso del término "troll" para describir el humor transgresor y subversivo se extendió por el grupo "hacktivista" Anonymus, que introdujo el término en un sitio web de Manga para aprovechar así sus posibilidades de difusión, al compartir con muchos otros en él sus denuncias de abusos. Lo que contribuyó a crear una más clara diferenciación entre el sentido (hasta entonces) más clásico del término, y el actual más moderno (Bishop, 2014).

Por su parte, los trollers amenazantes se denominan "snerts" en el ámbito de la investigación de trolls (Bishop, 2008, 2013), y son calificados erróneamente como trolls por los medios de comunicación. Aunque existe un acuerdo común en la presencia de un tipo de troller, conocido como "hater" (odiador). Se trata de un tipo muy específico de snert, que se esfuerza por intimidar a un objetivo específico (Bishop, 2012). Estos usuarios se dirigen a las víctimas de manera consciente, deliberada y no sienten la obligación de respetar al resto de personas. De hecho, algunos de ellos realizan acciones que pasan a tener un componente psicótico, por el cual sienten la necesidad de atacar a las personas que tienen más éxito que ellos (Bishop, 2013).

El troll, en el sentido clásico, puede ser visto como un entretenimiento consensuado de la comunidad con el fin de construir lazos entre los usuarios. Por el contrario, el nuevo fenómeno surgido denominado “hater”, se caracteriza por su discurso de odio en las redes sociales, donde se dedica a insultar sin ningún objetivo al resto de usuarios, ya sea por cuestión de raza, sexo, religión, política, etc.

### *3.3. La popularización de Internet y la Inteligencia Artificial*

Una de las características extrínseca del ciberespacio, que veremos con mayor detenimiento en el capítulo 3, es su popularización. Y, aunque en la actualidad, Internet se ha convertido en el método de interacción personal más popular, hemos visto que en sus orígenes no fue así. De este modo, en un principio la digitalización surgió como un objeto en el que únicamente cabía la lectura de contenidos, pasando a la acción con la interacción que fomentó la web 2.0.

Por su parte la web 3.0 es capaz de mejorar la gestión de los datos y apoyar la accesibilidad a Internet. Además, fomenta la creatividad, la innovación y la colaboración en las redes sociales. La web 3.0 es una web donde el concepto de sitio o página web desaparece, donde los datos no se poseen, sino que se comparten. Este cambio de estructura es peligroso, ya que los usuarios con pensamientos radicales pueden difundir su mensaje violento y de odio a todo el mundo mientras interactúan con otros de pensamientos similar, retroalimentándose continuamente.

Esta evolución ha continuado con la web 4.0, considerándola un sistema ultra inteligente, poderoso como el cerebro humano (Fowler & Rodd, 2017) . En otras palabras, las máquinas evolucionan de tal manera que pueden leer contenidos de la web y reaccionar ante ellos. Además, aunque son ideas en progreso y no hay una definición exacta, ya se habla de web 5.0 y web 6.0 (Raisoni & Sarode, 2016).

Así, la evolución de las nuevas tecnologías ha fomentado los avances de la inteligencia artificial, entendida como una entidad con la capacidad de observar y actuar como un ser humano. Como indica, Benko y Lányi (2009), puede ser un robot o un sistema software, dependiendo del entorno. También definida por Haenlein y Kaplan (2019) como "la capacidad de un sistema para interpretar correctamente los datos externos, para aprender de esos datos y utilizar ese aprendizaje para lograr objetivos y tareas concretas a través de la adaptación". Es decir, gracias a la inteligencia artificial podemos, entre otras cosas, gestionar grandes cantidades de datos a través de procesos de *big data* (Labrinidis & Jagadish, 2017) o de *machine learning* que nos permitirá el aprendizaje de las máquinas con la incorporación de nuevos datos (Mitchell et al., 2013). Como veremos en el siguiente capítulo, estas herramientas nos serán de gran utilizad en nuestra investigación para poder detectar los diferentes patrones de las características ambientas en la comunicación violenta y el discurso de odio en el ciberespacio.

En definitiva, Internet sigue en continua evolución convirtiéndose en una herramienta mundial para difundir pensamientos y opiniones entre sus usuarios. La inclusión de estas técnicas de inteligencia artificial ha generado un cambio en nuestros patrones de comportamientos, multiplicando los riesgos relacionados con el cibercrimen en general y la difusión de la comunicación y el discurso de odio en particular. Por ello, cada vez se hace más necesaria la regulación de este fenómeno y por supuesto, su prevención a través de la detección.

# Capítulo 2

# LA PREVENCIÓN (POR MEDIO DE LA DETECCIÓN) DE LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO EN INTERNET

"The technology alone is not enough. We also have to put the heart into it."

(Jane Goodall)

El aumento de la comunicación violenta y el discurso de odio en Internet plantea una serie de problemas que preocupan a la sociedad, por el sufrimiento y daño que puede causar a los usuarios y por los desórdenes sociales a los que puede conducir, más allá del ciberespacio (Ross et al., 2017; Waseem & Hovy, 2016). Podemos encontrar un ejemplo de este tipo de desórdenes recogido en el trabajo de Burnap y Williams (2016), en el que lo autores analizan el comportamiento de los usuarios de Twitter, tras el atentado terrorista del soldado Lee James Rigby en Woolwich (Londres), en el año 2013. Mostraron que dicho evento desencadenó una gran reacción social, tanto en Internet, en la que los mensajes de odio fueron evidentes dentro de las primeras horas después del suceso, como en el espacio físico, donde aumentaron los delitos e incidentes de odio. Para evitar este tipo de sucesos, es importante que se vigilen los mensajes de odio en Internet.

Para depurar responsabilidades, en el año 2017, las comisiones parlamentarias de Reino Unido y Alemania criticaron duramente a las principales redes sociales como

Twitter, Youtube (Google) y Facebook, alegando que no habían tomado las medidas necesarias ni habían sido suficientemente rápidas en la detección y eliminación del discurso de odio. De hecho, el gobierno alemán amenazó con sancionarlas con multas de hasta 50 millones de euros al año si continuaban así, dándoles como plazo máximo una semana (Thomasson, 2017). Cuando se les convocó a testificar ante el Comité de Asuntos Internos del Reino Unido, las empresas de medios sociales se negaron a revelar tanto el número de personas que empleaban para combatir el discurso de odio como la cantidad que gastaban en esto. No obstante, Google afirmó que había invertido "cientos de millones" y Facebook aseveró que tenía a miles de personas trabajando en esta cuestión (Gambäck & Sikdar, 2017).

Es indiscutible que los compromisos voluntarios de las empresas han dado lugar a mejoras. En este sentido, el 4 de febrero de 2019, la Comisión Europea publicó en un comunicado de prensa el avance que se había logrado en este ámbito. Ya que en el año 2016 cuando se puso en marcha el Código de conducta de la UE garantizando una rápida respuesta en la detección y retirada de discurso de odio en las redes sociales, únicamente evaluaban el 40% de los contenidos señalados en un plazo de 24 horas y retiraban el 28%, mientras que actualmente es de un 89% y 72% respectivamente. A pesar de estas buenas noticias, las empresas deben mejorar sus respuestas a los usuarios, aumentando la trasparencia en la información que se publica sobre la eliminación y el bloqueo del contenido ilegal en sus plataformas.

Pero, a pesar de los esfuerzos en la elaboración de normas y leyes por parte de los organismos y las inversiones millonarias que realizan las empresas de Internet para prevenir este tipo de comportamientos, recientemente se está comprobando que lo realmente eficaz para detectar y prevenir este tipo de comunicación, es el empleo de las técnicas de minería de datos automatizadas (Wei et al., 2017). Aunque el número de

investigaciones en este ámbito ha aumentado en los últimos años, su estudio sigue en una fase temprana (Fortuna & Nunes, 2018).

## 1. Prevención, Prohibición y Detección de la Comunicación Violenta y el Discurso de Odio en Internet (3 caras de la misma moneda)

La mayoría de los expertos y autoridades de muchos países comparten la opinión de que la comunicación violenta y, concretamente, el discurso de odio es un problema grave y debe prevenirse (Kaakinen et al., 2018). Sobre todo, a través de la educación y sensibilización de una sociedad en la que se eliminen los prejuicios y se puedan destruir los estereotipos. Y en la que exista una disuasión de la intolerancia, y un castigo de las formas más extremas y peligrosas del discurso de odio (Bhatnagar, 2018), cambiando así la opinión pública (Perry & Olsson, 2009).

Entre las medidas que propone Cohen-Almagor (2011) para contrarrestar el odio, se encuentran las basadas en la educación, tolerancia y respeto, con una mayor sensibilización y educación en el ciberespacio y en otras esferas de la sociedad, y el objetivo de conseguir mayor tolerancia y aceptación de la diversidad. Además, sugiere la publicación de los nombres de los sitios que incitan al odio, destacando su contenido, su ubicación y sus proveedores de servicio de Internet. Se pretende así alertar a los usuarios sobre esos lugares para incluirlos en programas de bloqueo, y enviarlos a los motores de búsqueda indexados para que los sitios puedan ser etiquetados adecuadamente.

Respecto a la cooperación internacional, el 4 de octubre de 2002, se fundó la Red Internacional Contra el Odio Cibernético (INACH). El objetivo de INACH era actuar colectivamente contra la discriminación, promoviendo la dignidad, el respeto y la

responsabilidad, permitiendo a los usuarios de Internet ejercer la libertad de expresión respetando al mismo tiempo los derechos y la dignidad de la sociedad. La INACH monitorea el ciberespacio y publica resúmenes e informes sobre la situación en diferentes países. Actualmente la organización está integrada por 22 países, que actúan como red coordinadora de las líneas directas especializadas en contenidos racistas y de odio.

Además, el Consejo de Europa apoyó una iniciativa llamada "No hate speech Movement" (Silva et al., 2016) y la UNESCO publicó un estudio titulado "Countering online hate speech", cuyo objetivo era ayudar a los países a hacer frente al problema (I Gagliardone et al., 2015).

En España, en la publicación por parte del Ministerio del Interior del Informe Raxen (2018), se propone la intervención mediante tres vías: el reproche social para que se evite la existencia de los seguidores de la intolerancia, el reproche por vía administrativa a través de sanciones económicas y el reproche penal en el que se generaliza el castigo a los delitos de odio. Siempre desde la educación, la tolerancia y los derechos humanos, como indica el artículo 10 de la Constitución Española. En este sentido, en diciembre de 2018, se lanzó el proyecto "Discurso de odio, racismo y xenofobia: mecanismos de alerta y respuesta coordinada (AL-RE-CO)", cuyo objetivo es aumentar las competencias de las autoridades del Estado para "identificar, analizar, monitorizar y evaluar el discurso de odio en las redes, a fin de diseñar estrategias compartidas frente al discurso motivado por racismo, xenofobia, islamofobia o antisemitismo". Entre los objetivos que se plantean desde el Ministerio del Interior (Informe Raxen, 2018), se encuentra que los miembros de las Fuerzas y Cuerpos de Seguridad adquieran una formación integral en el ámbito de los derechos humanos y los "delitos de odio", así como contrarrestar este fenómeno en las redes sociales. Para ello, las medidas que

proponen, entre otras, son las de colaborar en proyectos de investigación, y utilizar las nuevas tecnologías en el estudio del discurso de odio en las redes sociales. De este modo, se pretenden elaborar informes que posibiliten la identificación de patrones o tendencias en este tipo de delitos para que finalmente sean analizados y estudiados por los cuerpos policiales de manera objetiva. Siguiendo esta línea, el Ministerio del Interior aprobó la Instrucción 1/2019, 15 de enero de 2019 de la Secretaría de Estado de Seguridad, sobre el plan de acción de lucha contra los delitos de odio, cuyo objetivo es minimizar el daño producido por este tipo de delitos.

Para ello, el plan consta de cuatro pilares fundamentales. Por un lado, la formación de las Fuerzas y Cuerpos de Seguridad del Estado, la prevención de los delitos de odio, la atención a las víctimas y por último la respuesta eficaz y con rigor ante esta clase de sucesos (Ministerio del Interior., 2018).

Es cierto que las ONGs, los profesionales del Derecho, las empresas privadas e incluso la sociedad civil han desarrollado intervenciones, pero se sabe poco sobre su impacto (Blaya, 2019). En este sentido, el objetivo común es reducir los mensajes violentos y de odio, concretando estrategias para su prohibición, mediante la regulación. De este modo, para prevenir hay que prohibir y para regular hay que detectar. Así, tenemos tres caras de la misma moneda: Prevención, prohibición y detección.

## 2. La Regulación de la Comunicación Violenta y el Discurso de Odio en Internet

### *2.1. La Regulación en el ámbito público*

El debate sobre la regulación de la comunicación violenta y concretamente del discurso de odio en Internet, gira en torno a su realización sin vulnerar el derecho a la libertad de expresión para no generar una sensación de represión (Martínez et al., 2019).

En este sentido, el Pacto Internacional de Derechos Civiles y Políticos argumenta que, como derecho fundamental, la libertad de expresión debe primar, pero siempre y cuando se garantice la igualdad, la convivencia social, la paz, la dignidad humana y el derecho a vivir sin intimidación ni acoso (Cabo-Isasi & García-Juanatey, 2016).

Desde este punto de vista, el reconocimiento de la libertad de expresión como derecho individual permite que determinadas personas se amparen en esa libertad para propagar discursos a colectivos sociales concretos (Martínez et al., 2019). Por lo tanto, se plantea el reto de preservar esa libertad de expresión de las personas, al mismo tiempo que se protege la deshumanización, la incitación a la violencia y la discriminación (Blaya, 2019). No se puede obviar que la libertad de expresión es necesaria para mantener los derechos democráticos de los ciudadanos, facilitando el intercambio de opiniones entre ellos (Chetty & Alathur, 2018) pero siempre, y cuando, esas opiniones no causen un menoscabo a la integridad moral del resto.

Las leyes, reglamentos y otras medidas adoptadas por los Estados contra las páginas web y la propaganda difundida a través del ciberespacio, han tenido un efecto limitado, ya que los usuarios y administradores pueden estar en cualquier parte del mundo y, por lo tanto, son difíciles de localizar (Cohen-Almagor, 2011). La legislación y las políticas sobre la difusión del odio también varían de un país a otro. Por consiguiente, es evidente que se necesitan más y nuevas medidas para avanzar en la lucha contra este fenómeno.

En este sentido, Cohen-Almagor (2011) ya auguró lo que sucedería unos años después, indicando que se debía involucrar a las empresas privadas, especialmente a los proveedores de servicio de Internet (ISP), y a las empresas que proporcionan material para la difusión del contenido de odio. Además, propuso iniciativas en las que los usuarios de Internet solicitaban a los ISP que adoptaran un código de conducta

responsable, incluida una disposición contra el odio, e insistió en que los ISP aplicaran y respetaran sus propios códigos de conducta.

Asimismo, Cohen-Almagor (2011) hace referencia a la responsabilidad de los proveedores de servicios de Internet y las empresas de hospedaje de sitios web, encargadas de elaborar normas para prácticas responsables y aceptables de los usuarios. Porque, aunque anteriormente los gobiernos tenían la responsabilidad de gestionar la censura en el ciberespacio, en la actualidad esta actividad se subcontrata a proveedores comerciales de servicios de Internet, debido a que su mantenimiento resulta costoso y complicado de mantener en los presupuestos gubernamentales (Bieda & Halawi, 2015).

Esta reciente relación proporcionó a los gobiernos la flexibilidad que necesitaban para imponer prácticas de censura a los ISP, en lugar de obtener la aprobación de otras ramas del gobierno. Por ejemplo, el gobierno de Australia rechazó el filtrado de Internet y el administrador residente delegó a los ISP australianos la aplicación de la censura en la Red (Bieda & Halawi, 2015).

Además, Cohen-Almagor (2011) argumentó que los defensores de la lucha contra el discurso de odio debían explicar a los administradores de los ISP cuál era la naturaleza del odio, sus daños potenciales y por qué la responsabilidad corporativa hacia la comunidad implica la eliminación del contenido de sus servidores. Si conocen el fenómeno, es más probable que los proveedores de servicios de Internet tomen medidas proactivas para evitar la presencia de sitios que incitan al odio en sus servidores. De este modo, no sólo tomarían medidas una vez alertados, sino también antes, bloqueando y eliminando dichos sitios. Independientemente de cuáles sean los pasos que realicen las corporaciones para promover la seguridad en Internet, es fundamental que sean claros, concisos, transparentes y, sobre todo, razonados para los usuarios. Un ejemplo de cooperación entre una organización de monitoreo de Internet y un proveedor de

servicios, puede ser la Liga Antidifamación (ADL). YouTube, en diciembre de 2008, se puso en contacto con ADL por su experiencia en el tratamiento del odio en Internet. Gracias a dicha asociación, la ADL es ahora colaboradora del Centro de Abuso y Seguridad de YouTube, donde los usuarios tienen el poder de identificar y enfrentarse al odio, informando de los abusos que puedan observar (Cohen-Almagor, 2011).

Del mismo modo, en mayo de 2016, la Comisión Europea y varias de las principales empresas de Internet, anunciaron un nuevo "Código de conducta para luchar contra las declaraciones ilegales de odio *online*". Facebook, Twitter, Microsoft y YouTube acordaron especificar en sus condiciones de uso que se prohibía la incitación ilegal al odio. Además, acordaron establecer procesos claros y eficaces para revisar, eliminar o desactivar el acceso a dicho contenido en un plazo de 24 horas (Brown, 2018). En este sentido, en el Informe Raxen (2018) se indicó que se había alcanzado una tasa de eliminación del 70% del contenido ilegal, lo que demostró que se había logrado un progreso en este ámbito, uniéndose a este código de conducta otras plataformas, como Instagram, Snapchat y Daily motion.

En definitiva, la idea básica es que las empresas de Internet están mucho mejor situadas que los gobiernos para desarrollar e implementar términos de uso, códigos de conducta o normas comunitarias para los usuarios (Brown, 2018) . Además, no están obligadas a publicar o justificar sus acciones de eliminación o bloqueo de los usuarios que hayan infringido el código de conducta establecido, ya que no necesitan encontrar razones para hacerlo y justificarse ante un tribunal de justicia (Brown & Sinclair, 2019).

Es por ello, que todos los gobiernos nacionales deberían trabajar lo más estrechamente posible con las empresas de Internet para ayudar a encontrar formas de combatir la comunicación violenta y el discurso de odio en el ciberespacio (Brown, 2018).

### *2.2. La Regulación de Contenidos en las Redes Sociales*

Las redes sociales se han convertido en el medio preferido para expresar públicamente las ideas y pensamientos de los usuarios. Este hecho no excluye la necesidad de moderación cuya finalidad debe ser la de proteger a un usuario de otros violentos con diferentes ideas, eliminar ofensivas y presentar su mejor cara a sus anunciantes, nuevos usuarios y al público en general (Gillespie, 2018).

De entre las más destacadas, YouTube cuenta con sus "políticas"[2], Facebook con sus "normas comunitarias"[3] y Twitter con sus "reglas"[4], las cuales les permiten establecer el contenido permitido en la plataforma, dotándoles de cierta transparencia en lo que respecta a las decisiones de moderación.

Estas tres plataformas comparten estructura en la regulación de sus normas. Para empezar, justifican en el preámbulo la necesidad de la existencia de determinadas normas para conseguir un equilibrio entre la libertad de expresión y la protección. Seguidamente, estas normas se configuran en función de diferentes preocupaciones como son la violencia, el sexo, el discurso de odio, el acoso, las amenazas o la promoción del terrorismo entre otras. Por último, para conseguir que se cumplan las normas, cuentan con una combinación de herramientas tecnológicas para la detección automática de esta clase de contenidos, así como de revisores humanos que analizan las alertas automáticas y las denuncias externas formuladas por los propios usuarios u otros informadores, como pueden ser las ONGs o los organismos públicos.

[2] https://www.youtube.com/intl/es-419/about/policies/#community-guidelines
[3] https://www.facebook.com/communitystandards/
[4] https://help.twitter.com/es/rules-and-policies/twitter-rules

A pesar de que, en este sentido, estas plataformas no siempre han sido especialmente transparentes, en los últimos años han publicado informes periódicos que proporcionan algunos datos macro sobre el contenido principal contra el que han actuado (Miró-Llinares & Gómez-Bellvís, 2020)

### 2.2.1 *Las Políticas de YouTube*

Desde finales del año 2017, YouTube, recoge los datos globales relacionados con la cantidad de contenidos retirados en un informe trimestral de transparencia. Los últimos datos disponibles corresponden al periodo julio-septiembre de 2020. Durante estos meses se cancelaron 1.804.170 canales, retirándose los 32.900.267 vídeos que contenían. Para que se produzca esta cancelación, se deben dar tres escenarios: Que se registe un único caso de uso inadecuado grave; que se verifique que su único objetivo es transgredir las normas de YouTube o que el canal reciba tres faltas por incumplimiento de las normas de la comunidad en un periodo de 90 días.

El 85,4% de los canales retirados pertenecen a la categoría "spam, engaños o trampas"; el 6,5% a la "desnudos o sexual"; el 3% de los canales se retiraron por pertenecer a la categoría "inadecuado o que incita al odio"; el 2% a la categoría "protección infantil" y el 1,3% a la de "suplantación de identidad". Los canales correspondientes a las categorías "acoso y ciberacoso" (0,8%) y "promoción de la violencia y el extremismo violento" (0,05%) no alcanzan el 1%.

El 93,88% de los vídeos retirados se detectaron de manera automática. La mayor parte de ellos corresponden a las categorías "protección infantil" (31,7%), "spam, engaño o trampas" (25,5%) y "contenido con desnudo o sexual" (20%). El 14,2% de los vídeos retirados pertenecen a la categoría "violento o gráfico", el 2,5% a la "promoción de la violencia y el extremismo violento" y el 1,1% estaban relacionados con el contenido "inadecuado o que incita al odio".

De igual manera, el 99,6% de los comentarios se detectaron de manera automática, catalogándose el 51,4% como spam. El 26,9 % pertenecen a la categoría "protección infantil", l 17,3 % al "acoso y ciberacoso" y el 4,1% se refería a contenidos inapropiados o de incitación al odio. En lo referente a las denuncias, durante este trimestre, se denunciaron más de 15.000.000 de vídeos. El 99,6% las realizaron los usuarios por diversas razones. Las denuncias por contenido engañoso o spam fueron las más habituales junto con las de contenido sexual, 27% y 20% respectivamente. El 18,7% fueron por contenido abusivo o incitación al odio y el 13,2% por contenido violento o repulsivo. El resto, con menor incidencia, fueron por actividades peligrosas o engañosas (8,4%), maltrato infantil (6,3%) y por fomentar el terrorismo (5,9%).

### 2.2.2 *Las Normas Comunitarias de Facebook*

Posiblemente, la política de contenidos de Facebook sea la más definida ya que cuenta con un sistema de regulación más desarrollado. Quizás, se deba a la partición en diversas polémicas (Gillespie, 2018) o que realmente es consciente de lo importante que es esta cuestión para que la red social sigua siendo la que es. Facebook establece cuatro valores de protección en su contenido: autenticidad, seguridad, privacidad y dignidad.

Además, divide sus normas comunitarias en seis apartados. Tres de ellos regulan aspectos relacionados con la comunicación violenta y el discurso de odio: violencia y comportamiento criminal, seguridad y contenido cuestionable. Estos mismos apartados contienen subapartados que establecen restricciones relacionadas con nuestra investigación: (1) violencia e incitación; (3) Organización de actividades nocivas y publicidad de la delincuencia y (12) discurso de odio o incitación al odio. En este último, Facebook define el discurso de odio como "el ataque directo a personas basado en características protegidas, como puede ser la raza, etnia, nacionalidad, religión, clase,

orientación sexual, sexo, identidad sexual y la discapacidad o enfermedad grave". En este sentido, divide los ataques en tres categorías:

- Lenguaje violento o que se refiere a una persona o grupo de personas relacionándolos con crímenes, animales, insectos, bacterias o enfermedades. Incluye también utilizar estereotipos que se consideran deshumanizadores,
- Generalizaciones que impliquen inferioridad, insultos y expresiones de desprecio, rechazo o repulsión.
- Llamamientos a la segregación, la exclusión en general y la exclusión política, económica y social.

Además, al igual que YouTube, Facebook publica desde el año 2017 informes trimestrales en los que la información es mucho más completa que en el resto de redes sociales. Se estructuran en relación con los grupos de normas comunitarias, donde se indica el número de elementos retirados, su prevalencia, la proporción de contenido que fue detectado de manera automática, la cantidad de veces que la retirada fue apelada y el porcentaje de veces que se aceptó la apelación. Los datos relacionados con el último trimestre disponible comprenden los meses de julio, agosto y septiembre de 2020. Como novedad, presentan datos de la prevalencia del discurso de odio en Facebook. En este sentido, la tasa de contenido analizado y proactivo para el discurso de odio se ha mantenido prácticamente igual en el segundo y tercer trimestre de 2020. Su prevalencia se encuentra entre el 0,10% y el 0,11% de las visitas en el tercer trimestre. En este trimestre, se retiraron 22,1 millones de elementos relacionados con el discurso de odio. El 94,7% del contenido retirado fue identificado por Facebook antes de ser denunciado por los usuarios, aumentando en gran proporción respecto al 25% que se detectó en 2017. En este periodo únicamente se apelaron 41.000 piezas de contenido, cifra muy inferior a la registrada en el primer trimestre de 2020 (1.200.000). De todas las

apelaciones únicamente se restauraron 4.700 elementos, el resto (232.000) fueron restaurados sin apelación.

### 2.2.3 *Las Reglas de Twitter*

La red social Twitter enumera diferentes reglas en función de las cuales puede eliminar ciertos contenidos que han publicado sus usuarios. Al igual que en Facebook, estas reglas se dividen en diversos apartados, aunque aquí únicamente se indican tres: Seguridad, privacidad y autenticidad. A diferencia de Facebook, Twitter incluye la dignidad dentro de la seguridad. En el apartado seguridad encontramos las reglas relacionadas con la comunicación violenta y el discurso de odio, concretamente en los siguientes puntos, similares a los de Facebook: violencia, terrorismo/extremismo violento, abuso/acoso y comportamiento de incitación al odio. En el apartado relativo a la violencia incluye, por un lado, discursos que tienen una relación directa con la violencia física, amenazas violentas y por el otro, los que tienen una relación indirecta como puede ser la glorificación de esa violencia. Además, incluye un apartado sobre la incitación al odio con similares características a las normas comunitarias de Facebook.

Twitter se demoró un año más que las anteriores en publicar informes sobre las denuncias contra los usuarios por violar sus reglas. El último informe disponible es el relacionado con el segundo semestre de 2019. En él se indica que se han procesado 2.316.314 de cuentas, siendo un 47% en comparación con el último periodo de informe. Del mismo modo, en este último semestre se han suspendido 872.855 cuentas, aumentando un 27% en comparación con el último periodo de informe. Y se ha eliminado 2.863.181 de contenido, aumentando en un 50% en relación con el último periodo del informe. De esta cantidad, 1.445.181 era contenido de odio.

De este modo, observamos que Twitter está activamente en un proceso en curso para que sus usuarios cumplan las nuevas directrices que han estipulado relacionadas con la comunicación violenta y el discurso de odio (Sharma et al., 2018). Como ya se ha indicado, las redes sociales no sólo quieren identificar a los usuarios que amenazan con violencia o daño físico, también quieren buscar cuentas relacionadas con grupos que promuevan esa violencia contra otros usuarios. En este sentido, cualquier contenido que glorifique la violencia o a los autores de un hecho violento, estará quebrantando las nuevas directrices de Twitter para combatir el discurso de odio.

Una vez analizas las tres plataformas, vemos que existen ciertas diferencias en la redacción y articulación de sus normas, pero el objetivo y las conductas prohibidas son prácticamente similares. Todas ellas mencionan la necesidad del contexto para analizar la comunicación violenta y el discurso de odio y la importancia que tiene su detección para poder prevenir y regular estas conductas.

## 3. La Detección Automática de la Comunicación Violenta y el Discurso de Odio

Según Fortuna y Nunes (2018), existen diferentes razones para centrarse en la detección automática de la comunicación violenta y el discurso de odio. En primer lugar, debemos recordar las diferentes iniciativas llevabas a cabo por la Comisión de la Unión Europea para reducir la comunicación violenta y en especial, el discurso de odio. Como, por ejemplo, la creación de varios programas para la lucha contra la incitación al odio o la presión a las redes sociales YouTube, Twitter, Facebook, y Microsoft para que firmaran un código de voz de la UE, que incluía el requisito fundamental de revisar la mayoría de las notificaciones válidas para la eliminación de la incitación al odio ilegal.

Por otro lado, no se disponen de técnicas automáticas y esto es fundamental, porque tienen como objetivo clasificar de manera automática el texto como discurso de odio,

haciendo que su detección sea más fácil y rápida (Martins et al., 2018) para aquellos que tienen la responsabilidad de proteger al público, tanto por parte de la policía como de los proveedores de servicios (Burnap & Williams, 2016; Ross et al., 2017). En este sentido, Schmidt y Wiegand (2017) ofrecen una visión general breve, estructurada, completa y crítica del campo de la detección automática del discurso de odio en el procesamiento del lenguaje natural. Su investigación se divide en varias secciones, presentando en primer lugar la terminología necesaria para estudiar el fenómeno. Posteriormente, analizan con mayor profundidad las características utilizas en esta problemática y se centran en la investigación sobre la intimidación. También presentan un apartado dedicado a los métodos de clasificación, los desafíos y otra sobre datos.

De manera complementaria, la investigación realizada por Fortuna y Nunes (2018), proporciona una definición más detalla de la problemática analizada, compara la incitación al odio con otros conceptos relacionados, subtipos de odio y enumera reglas que son útiles en la tarea de la clasificación del discurso de odio. Además, utilizan un método sistemático y analizan no solo documentos que se centran en algoritmos, sino también en estadísticas descriptivas sobre la detección del discurso de odio, ofreciendo una visión general de la evolución del área en los últimos años.

Otra de las razones para centrarse en la detección de la comunicación violenta y el discurso de odio en Internet, es que no existen datos suficientes, Es decir, se carece de monitorización automática, documentación y recolección de datos sobre el odio y la violencia contra las personas lesbianas, gays, bisexuales, transgénero e intersexuales (ILGA, 2016). Sin embargo, la detección del discurso de odio es una tarea muy importante porque, como entre otras cosas, se relaciona con los delitos de odio en el espacio físico (Ross et al., 2017; Waseem & Hovy, 2016) y su análisis podría aportar información de utilidad sobre este fenómeno.

Igualmente, algunas compañías y plataformas podrían estar interesadas en la detección y eliminación del discurso de odio en la Red (Wendling, 2015). Por ejemplo, los editores de medios de comunicación y plataformas *online*, en general, necesitan atraer a los anunciantes y, por tanto, no pueden arriesgarse a ser conocidos como plataformas en las que se fomente o se permita la incitación al odio (Hewitt el al., S., 2016). Además, los usuarios de estas plataformas podrían bloquear esos comentarios con expresiones también de odio para evitar exponerse a ellas, siendo contraproducente (Fortuna & Nunes, 2018).

Por último, la calidad del servicio debe ser la adecuada. Es decir, las empresas de medios sociales en Internet tienen como objetivo fundamental facilitar la comunicación y el respeto entre sus usuarios (Oboler, A., & Connelly, 2014). Para conseguirlo, se deben adoptar medidas que desanimen a los usuarios a que publiquen mensajes discriminatorios y en el caso de no conseguirlo, se eliminen en un plazo razonable de tiempo.

### *3.1 ¿Cómo se Detecta la Comunicación Violenta y el Discurso de Odio en Internet?*

El objetivo de las técnicas automatizadas es clasificar el texto como comunicación violenta o discurso de odio, facilitando su detección (Martins et al., 2018). Una forma de detectar el discurso de odio, es utilizar un enfoque basado en el léxico, tal como lo presenta Gitari y sus colaboradores (2015) . Mientras tanto, como afirman otras investigaciones (Davidson et al, 2017), los métodos de detección léxica tienden a tener baja precisión porque clasifican todos los mensajes que contienen términos particulares como expresiones de odio. En este apartado se realizará una revisión de las investigaciones más relevantes en la detección del discurso de odio para conocer las técnicas que se están aplicando.

### 3.1.1 *Algoritmos para la detección de la comunicación violenta y el discurso de odio*

En los últimos años han aumentado las investigaciones científicas cuyo objetivo era la detección del discurso de odio en Internet. En este sentido, ser realizó una búsqueda por palabras clave en el repositorio ProQuest que arrojó 1.055 resultados hasta el año 2000, mientras que en el periodo del 2001 al 2010 se llegó a la cifra de 5.081. Desde el año 2010 hasta el 2020, ya existen 6.831 publicaciones, lo que supone un incremento de más del 500%.

En lo que se refiere a la metodología, los investigadores utilizan técnicas de *machine learning* para la clasificación de este tipo de lenguaje. El *machine learning* o aprendizaje automático consiste en el aprendizaje de un software a través del ajuste de determinados algoritmos en función de los datos aportados a su sistema. Es decir, el objetivo de esta técnica es crear sistemas que "aprendan" de manera automática, identificando patrones en gran cantidad de datos (Alpaydin, 2020).

Por ejemplo, el Random Forest, es una de las técnicas supervisadas de *machine learning* que combina un número *K* de árboles de decisión para conocer las variables ambientales que más influyen sobre la variable dependiente (Breiman, 2001; Rodriguez-Galiano et al., 2015). Los árboles de decisión, conocidos también como CART (Breiman et al., 1984), son técnicas que se utilizan con fines descriptivos y predictivos. Esta técnica trata de generar de forma gráfica y analítica, en forma de árbol, todos los comportamientos que se pueden dar a partir de una acción.

### 3.1.2 *Enfoques de text mining para la detección automática de la comunicación violenta y el discurso de odio*

El *text mining* o minería de textos tiene por objeto revelar la información oculta mediante métodos que, por un lado, son capaces de hacer frente al gran número de palabras y las estructuras en el lenguaje natural y, por el otro permiten manejar la vaguedad, la incertidumbre y la confusión que generan, comprendiendo el conjunto de ese texto. Hotho y sus colaboradores (2005) describen la minería de textos como un método interdisciplinario sobre la recuperación de información, el aprendizaje automático, la estadística, la lingüística computacional y especialmente la minería de datos.

La minería de textos o el descubrimiento de conocimiento a partir de textos (KDT, por sus siglas en inglés), consiste en el análisis de texto apoyado por la máquina. Utiliza técnicas de recuperación y extracción de información, así como el procesamiento del lenguaje natural (PNL) y las conecta con los algoritmos y métodos del descubrimiento del conocimiento a partir de datos (KDD, por sus siglas en inglés), la minería de datos, el aprendizaje automático y la estadística.

Para enriquecer la definición de la minería de textos, podemos hacer referencia a las distintas áreas de investigación relacionadas. Para cada una de ellas, podemos dar una definición diferente, motivada por la perspectiva específica del área a investigar (Hotho et al., 2005):

- Minería de textos como extracción de información. En este enfoque se asume que la minería de textos corresponde principalmente a la extracción de información, como la extracción de hechos de los textos.

- Minería de textos como text data mining. La minería de textos puede definirse de manera similar a la minería de datos o data mining, como la aplicación de algoritmos y métodos computacionales a los textos con el objetivo de encontrar patrones útiles. El mayor problema de este enfoque es que es necesario preparar los documentos mediante su preprocesamiento. El text data mining utiliza métodos de extracción de información o algunas técnicas más sencillas de preprocesamiento para extraer datos de los textos. A estos datos extraídos ya se le podrían aplicar algoritmos de minería de datos (Gaizauskas, 2002).

En nuestra investigación, consideramos que la minería de textos es principalmente una minería de datos de textos. Por lo tanto, nuestro enfoque se basará en los métodos que extraen patrones útiles de los textos para categorizar o estructurar colecciones de textos o para extraer información útil de ellos.

En este sentido, encontrar las herramientas más adecuadas para solucionar un problema de clasificación puede ser una de las tareas más exigentes cuando se utiliza la minería de datos dentro del aprendizaje automático. Por ello, vamos a describir las técnicas que han utilizado los investigadores más relevantes en este ámbito.

#### 3.1.2.1 Técnicas de text mining

La mayoría de las investigaciones en el campo de la detección automática del discurso de odio tratan de adaptar las estrategias ya conocidas de minería de textos a dicho problema. A continuación, se presenta una breve revisión de las técnicas comúnmente utilizadas en la minería de textos, comenzando por los enfoques más

sencillos que utilizan diccionarios léxicos hasta los más complejos como el análisis de sentimientos o el *deep learning*:

##### 3.1.2.1.1 Diccionarios de palabras

Una de las estrategias más habituales en la minería de textos es el uso de diccionarios. Este enfoque consiste en compilar una lista de palabras relevantes (i.e., el diccionario) que posteriormente se buscan y se cuentan en el texto. En el caso de la detección del discurso de odio, se ha llevado a cabo utilizando:

(1) palabras contenidas en el texto, tales como insultos y blasfemias y pronombres personales, recogidas en repositorios como Hatebase (https://hatebase.org/) o Swear Word List & Curse Filter (https://www.noswearing.com/) (Liu y Forss, 2015); (2) número de palabras profanas en el texto, con un diccionario que consta de 414 palabras, incluyendo acrónimos y abreviaturas, donde la mayoría son adjetivos y sustantivos (Dadvar et al., 2012); (3) el léxico de la ortografía para la detección de afecto negativo, ya que contiene una lista de palabras que denotan una connotación negativa y puede ser útil, porque no todos los comentarios groseros necesariamente contienen blasfemias y pueden ser igualmente dañinos (Dinakar et al.,2011).

##### 3.1.2.1.2 Distancias métricas

Algunos estudios (Nobata et al., 2016; Warner & Hirschberg, 2012) han indicado la posibilidad de que las palabras ofensivas de los mensajes de texto se oculten a través de un error ortográfico intencionado. En la mayoría de los casos, simplemente con la sustitución de un solo carácter, quizás para eludir los filtros empleados por las webs. Ejemplos de estos términos son "@ss", "sh1t" (Nobata et al., 2016), "nagger" (Warner & Hirschberg, 2012). En este sentido, la distancia Levenshtein (o distancia entre palabras), entendida como el número mínimo que se necesita para transformar una

palabra en otra, puede utilizarse con este fin, como muestran (Nandhini & Sheeba, 2015) en la investigación que desarrollaron sobre el ciberacoso en las redes sociales. Los autores colaboraros con el gobierno para que tomaran medidas antes de que muchos usuarios se convirtieran en víctimas de este fenómeno. La distancia métrica se puede utilizar para completar los enfoques basados en diccionarios.

##### 3.1.2.1.3 Bolsas de palabras (Bag-Of-Words, BOW)

Otro modelo similar a los diccionarios es la bolsa de palabras (Burnap & Williams, 2016; Greevy & Smeaton, 2004; Kwok & Wang, 2013). En este caso, se crea un corpus basado en las palabras que están en el texto a analizar, en lugar de un conjunto predefinido de palabras, como en los diccionarios. Después de recoger todas las palabras, la frecuencia de cada una de ellas se utiliza como característica para entrenar a un clasificador. Las desventajas de este dipo de enfoques son que se ignora la secuencia de palabras, así como su contenido sintáctico y semántico. Por lo tanto, puede llevar a una clasificación errónea si las palabras se utilizan en contextos diferentes. Para superar esta limitación se pueden adoptar N-gramas.

##### 3.1.2.1.4 N-gramas

Los N-gramas es una de las técnicas más utilizadas en la detección automática de la comunicación violenta y concretamente del discurso de odio (Badjatiya et al., 2017; Burnap & Williams, 2016; Davidson et al., 2017; Greevy & Smeaton, 2004), así como de otras cuestiones relacionadas con este fenómeno (Liu & Forss, 2015; Nobata et al, 2016; Waseem & Hovy, 2016) El enfoque más común de los n-gramas consiste en combinar palabras secuenciales en listas con el tamaño *n*. Es decir, a partir de una palabra se obtienen subcadenas de *n* caracteres. Por lo tanto, por cada una de las palabras que se encuentran en el documento se proporcionará una representación distribuida de las mismas en un conjunto de n-gramas (Cavnar, 1995). Esto permite

mejorar el desempeño de los clasificadores, ya que incorpora en cierto grado el contexto en cada palabra. En lugar de utilizar palabras, también es posible utilizar n-gramas con caracteres o sílabas. Este enfoque no es tan susceptible a las variaciones ortográficas como cuando se usan las palabras. En función de los *n*, denominaremos bigramas (o digramas) si n=2 y trigramas si n=3 y así, sucesivamente.

Existen investigaciones como la de Mehdad y Tretreault (2016) respecto al lenguaje abusivo, en las que gracias a la técnica del N-grama se obtuvieron buenos resultados. Sin embargo, su uso también tiene desventajas. Una de ellas es que las palabras relacionadas pueden tener una gran distancia en una oración (Burnap & Williams, 2016) y una solución para este problema, sería la de aumentar el valor *n*, pero ralentizaría la velocidad de procesamiento (Chen, 2011). Además, los estudios señalan que los valores de *n* más altos (Banerjee et al., 2012) se comportan mejor que los valores más bajos (unigramas y trigramas) (Liu & Forss, 2014).

En definitiva, las características de los n-gramas a menudo se presentan como altamente predictivas en el problema de la detección automática de la comunicación violenta y el discurso de odio, pero funcionan mejor cuando se combinan con otras técnicas (Schmidt & Wiegand, 2017).

##### 3.1.2.1.5 TF-IDF (Term Frecuency-Inverse Document Frecuency)

El “Term Frequency” (TF) hace referencia a la cantidad de veces que una palabra aparece en un documento. Así, el “Inverse Document Frequency” (IDF) se refiere a la puntuación que le corresponde en función de lo extraña que es la palabra en los documentos, expresándose con el logaritmo de la frecuencia inversa del término o palabra en todo el documento. Por lo tanto, el TF-IDF es una medida de la importancia de una palabra en un documento dentro de un corpus que aumenta en proporción al número de veces que esa palabra aparece en el mismo (Dinakar et al., 2011). Sin

embargo, se distingue de una bolsa de palabras, o N-gramas, porque la frecuencia del término se compensa con la frecuencia de la palabra en el corpus, lo que compensa el hecho de que algunas palabras aparecen con más frecuencia en general (Fortuna & Nunes, 2018). Como por ejemplo las palabras vacías o *stop words*, que se suceden frecuentemente en el texto pero que no aportan significado relevante, como pueden ser las preposiciones, las conjunciones y los artículos.

A diferencia de la mayoría de los estudios hasta el momento, en los que únicamente se consideraba la frecuencia o el recuento de frecuencia de N-gramas, Liu y Forss (2014) emplearon la técnica de N-gramas ponderados por TF-IDF en la construcción de los modelos de clasificación del contenido. Sus resultados mostraron que los modelos basados en unigramas, aunque eran mucho más simples, mostraban su valor y eficacia únicos en la clasificación del contenido en la web. Del mismo modo, otras investigaciones sobre la detección del discurso de odio (Badjatiya et al., 2017) concluyen que, entre los métodos empleados el TF-IDF es mejor que el de N-gramas.

##### 3.1.2.1.6 POS (Part Of Speech)

Los enfoques de "Part Of Speech" o POS, permiten mejorar la importancia del ambiente y detectar el papel de la palabra en el contexto de una frase. Estos enfoques consisten en detectar cuál es la categoría de la palabra; por ejemplo, pronombre personal (PRP), verbo 3ª persona singular presente (VBP), Adjetivos (JJ), Determinantes (DT), formas base verbales (VB), etc.(Fortuna & Nunes, 2018).

La técnica POS también se ha utilizado para detectar el lenguaje dañino y discriminador, junto con la bolsa de palabras y los bigramas (Greevy & Smeaton, 2004). Por ejemplo, en la investigación realizada por Dinakar y sus colaboradores (2011), fue posible identificar pares de bigramas frecuentes, como PRP_VBP, JJ_DT y VB_PRP, que se mapearían como "tú eres". También se usó esta técnica en la investigación de

Burnap y Williams (2014) para detectar el discurso de odio racial, por etnia o religión en Twitter, identificando frases como “enviarlos a casa” o “deberían ser colgados”. Sin embargo, la muestra que carecía de expresiones de odio también mostró una abundancia de patrones similares, como “déjalos en paz” o “son pacíficos”.

##### 3.1.2.1.7 LSF (Lexical Syntactic Feature-based)

En el estudio realizado por Chen (2011) se utilizó el procesador de lenguaje natural, propuesto por el *Stanford Natural Language Processing Group*, para detectar el contenido ofensivo e identificar a los potenciales usuarios dañinos en las redes sociales. Los rasgos obtenidos fueron pares de palabras en la forma “(gobernador, dependiente)”, donde el dependiente es un aposicional del gobernador (por ejemplo, “Tú, por supuesto, un idiota” significa que “idiota”, el dependiente, es un modificador del pronombre “tú”, el gobernador). En particular, incorporaron el estilo de escritura de los usuarios, la estructura y el contenido específico del ciberacoso como características para predecir la potencialidad de los usuarios para enviar contenido ofensivo Los resultados de los experimentos realizados mostraron que el empleo de la LSF logró un rendimiento significativamente mejor que los métodos existentes en la detección del contenido ofensivo.

##### 3.1.2.1.8 Dependencias tipográficas

La representación de las dependencias mecanografiadas de Stanford se diseñó para proporcionar una descripción sencilla de las relaciones gramaticales en una frase que puede ser fácilmente comprendida y utilizada eficazmente por personas sin conocimientos lingüísticos que quieran extraer relaciones textuales (De Marneffe & D. Manning, 2008). La representación de las dependencias mecanografiadas se ha empleado para extraer patrones gramaticales temáticos como muestra la investigación desarrollada por Gitari y sus colaboradores (2015) sobre la detección del discurso de

odio, así como en el desarrollo del clasificador de aprendizaje automático supervisado para el contenido de odio en Twitter de Burnap y Williams (2014, 2015a), reduciendo los falsos negativos en un 7% sobre la "bolsa de palabras".

##### 3.1.2.1.9 Clasificación en temas

Esta técnica tiene como objetivo descubrir el tema abstracto que aparece en un documento. Agarwal y Sureka (2017) realizaron un estudio en el sitio web de microblogging Tumblr, en el que utilizaron características lingüísticas de modelado de temas para identificar los mensajes racistas o radicalizados. De este modo, entrenaron el modelo identificando varios rasgos semánticos, de sentimiento y lingüísticos del texto de forma libre. Los resultados experimentales mostraron que el enfoque propuesto era efectivo.

##### 3.1.2.1.10 Análisis del sentimiento

Los recursos léxicos se utilizan a menudo para buscar palabras negativas específicas en los mensajes (como insultos, difamaciones, etc.), ya que la presencia de estas palabras puede ser una característica predictiva de la expresión de odio (Wei et al., 2017). Teniendo en cuenta que este tipo de lenguaje tiene una polaridad negativa, los autores han empelado el cálculo del sentimiento como técnica para la detección del discurso de odio (Agarwal & Sureka, 2017; Davidson et al., 2017; Del Vigna et al., 2017; Gitari et al., 2015; Liu & Forss, 2014). Debido a esto, varios enfoques reconocen la relación entre el odio y el análisis de sentimientos al incorporar este último como un auxiliar para llevar a cabo la clasificación (Schmidt & Wiegand, 2017). Dinakar et al. (2012), Sood et al. (2012) y Gitari et al. (2015) siguen un enfoque de varios pasos, en el que se aplica un clasificador dedicado a detectar la polaridad negativa antes de que el clasificador especializado verifique la evidencia de la incitación al odio. Además, Gitari et al. (2015) ejecutaron un clasificador adicional que eliminaba las frases no subjetivas

antes de la polaridad antes mencionada. Además de los enfoques de varios pasos, también existen enfoques de un solo paso que incluyen alguna forma de información sobre los sentimientos como característica. Por ejemplo, Van Hee y sus colaboradores (2015) utilizaron el número de palabras positivas, negativas y neutras (según un léxico de sentimientos) que aparecen en un texto de comentario dado. Otros intentos de aislar el subconjunto de la incitación al odio del conjunto de expresiones polares negativas se basan en la observación de que la incitación al odio también muestra un alto grado de polaridad negativa (Burnap et al., 2013; Sood & Churchill, 2012). Algunos autores como Liu y Forss (2014) suelen utilizar esta técnica en combinación con otras, como por ejemplo n-gramas obteniendo mejoras significativas en los niveles de precisión para la web analizada relacionada con la violencia y el racismo.

##### 3.1.2.1.11 Deep Learning

El *deep learning* o el aprendizaje profundo es una rama del *machine learning* o aprendizaje automático, en el que un conjunto de algoritmos hace uso de redes neuronales para aprender a partir de la experiencia. Es imprescindible que estos algoritmos de aprendizaje automático "entrenen" a partir de ejemplos que ya existen. Las técnicas de aprendizaje profundo también se están utilizando recientemente en la clasificación de textos y el análisis de sentimientos, con una gran precisión (Gambäck & Sikdar, 2017; Pitsilis et al., 2018; Yuan et al., 2016).

En este sentido, la investigación desarrollada por Yuan y sus colaboradores (2016) tuvo como objetivo el descubrimiento de la discriminación de los tweets utilizando este tipo de modelos. Desarrollaron un modelo de aprendizaje profundo de dos fases, en el que aprendió representaciones de texto basadas en tweets, débilmente etiquetados, con algunos hashtags específicos y posteriormente entrenaron al clasificador en un conjunto pequeño de datos bien etiquetados. Los resultados experimentales muestran que el

método propuesto se puede utilizar de manera satisfactoria para la identificación de la discriminación. Del mismo modo, Pitsilis y sus colaboradores (2018) también abordaron la problemática de discernir el contenido de odio en los medios sociales, relacionados con el racismo y el sexismo a través del lenguaje profundo. Estos autores afirman que los modelos de *deep learning* tienen un alto potencial para clasificar textos o analizar el sentimiento general.

#### *3.1.2.2 Técnicas complementarias para la detección de la comunicación violenta y el discurso de odio*

Como complemento a los enfoques comúnmente utilizados en el análisis de la minería de textos, se están utilizando varias técnicas específicas para abordar el problema de la detección automática de la comunicación violenta y el discurso de odio (Fortuna & Nunes, 2018).

##### 3.1.2.2.1 Características del ofensor

Algunos otros estudios también consideran características más relacionadas con el gráfico de la red social. En este caso en particular, la investigación sobre el discurso de odio en Twitter de Waseem y Hovy (2016), vinculaba los mensajes disponibles del mismo usuario y se centraba en las características del usuario como el género y la localización geográfica.

##### 3.1.2.2.2 Declaraciones de superioridad del grupo

Además de la cuestión de la objetividad y la subjetividad del lenguaje, las declaraciones de superioridad del grupo también pueden considerarse expresiones de odio. En este caso, la incitación al odio también puede estar presente cuando sólo hay declaraciones defensivas o de orgullo, en lugar de ataques dirigidos a un grupo específico como argumentan Warner y Hirchberg (2012) en su investigación.

3.1.2.2.3 Centrarse en estereotipos particulares

En algunos estudios (Warner & Hirschberg, 2012) los autores plantean la hipótesis de que la incitación al odio a menudo emplea estereotipos bien conocidos y, por lo tanto, subdividen dicha expresión de acuerdo con los siguientes criterios estereotipos. Este enfoque puede ser útil, porque cada estereotipo tiene un lenguaje específico: palabras, frases, metáforas y conceptos. Por ejemplo, en el estudio realizado por Warner y Hirschberg (2012) el discurso antihispánico podría hacer referencia al cruce de fronteras; el discurso antiafricano-americano a menudo hace referencia al desempleo o a la educación de padres solteros; y el lenguaje antisemita a menudo se refiere al dinero, la banca y los medios de comunicación. De este modo, como indican los autores, la creación de un modelo de lenguaje para cada estereotipo es un requisito previo necesario para construir un modelo general para todas las expresiones de odio.

3.1.2.2.4 Interseccionismo de opresión

Como indican Fortuna y Nunes (2018), la interseccionalidad es un concepto que señala la conexión entre varios tipos particulares de expresiones de odio (p. ej., la prohibición del burka puede analizarse como una conducta islamofóbica o sexista, ya que este símbolo es utilizado por los musulmanes, pero sólo por las mujeres). El interseccionismo de varios tipos de opresiones presenta un desafío particular para la identificación automatizada de las expresiones de odio y por ello ha sido tenido en cuenta en la literatura. En el estudio de Burnap y Williams (2016), la intersección de los subtipos de odio es considerada sólo en la evaluación del modelo, donde más de una clase fue considerada al mismo tiempo.

## *3.2 Recapitulación*

Estas son algunas de las técnicas automatizadas más importantes en la lucha contra la comunicación violenta y el discurso de odio en Internet, máxime cuando nos

encontramos en una sociedad en la que los medios *online* juegan un papel significativo en la interacción entre las personas. Por ello, es positivo conocer que las investigaciones están aumentando en los últimos años. Warner y Hirschberg (2012) y Burnap y Williams (2015a) han sido de los primeros investigadores en utilizar clasificadores basados en el aprendizaje automático para detectar el lenguaje abusivo. Djuric et al., (2015) y Mikolov (2013) han incorporado la representación de palabras incrustadas y Nobata et al. (2016) ha combinado elementos de lenguaje predefinidos e incrustaciones de palabras para entrenar un modelo de regresión. Waseem (2016) usó regresión logística con N-gramas y características específicas de usuario como género y ubicación. Davidson et al., (2017) llevaron a cabo una investigación más profunda sobre diferentes tipos de lenguaje abusivo. Badjatiya et al., (2017) experimentaron con modelos basados en el aprendizaje profundo utilizando clasificadores de impulso de gradiente en conjunto para realizar una clasificación multiclase sobre el lenguaje sexista y racista, entre otras muchas investigaciones que hemos citado. Sin embargo, todavía quedan algunos desafíos pendientes.

## 4. Limitaciones en los Enfoques Tradicionales para la Detección de la Comunicación Violenta y el Discurso de Odio en Internet

La detección de la comunicación violenta y el discurso de odio requiere conocimiento sobre la cultura y la estructura social, ya que el lenguaje evoluciona rápidamente, sobre todo entre los jóvenes que se comunican a través de las redes sociales (Raisi & Huang, 2016). Por otro lado, como indican Kwok y Wang (2013), si para los seres humanos es complicado definir estos conceptos y su clasificación, todavía lo es más para las máquinas, porque a pesar de la naturaleza de la incitación al odio, el lenguaje abusivo puede ser muy fluido y gramaticalmente correcto, puede cruzar los

límites de la frase, siendo común la utilización del sarcasmo. Para Nobata y sus colaboradores (2016) , la evolución de los fenómenos sociales y del lenguaje dificulta el seguimiento de todos los insultos raciales y de las minorías, por lo tanto, la detección del discurso del odio es más que una simple búsqueda de palabras clave. Además, como indican Fortuna y Nunes (2018), no existen conjuntos de datos comunes, ni siquiera para la investigación en inglés. Por lo tanto, la definición de un set de datos principal es fundamental. Este sería un paso importante para facilitar la comparación entre los diferentes estudios.

En las investigaciones que hemos mencionado, se ha identificado la descripción de los métodos, las características extraídas y los algoritmos utilizados. Sin embargo, es raro encontrar trabajos con código abierto. De este modo, un mayor intercambio de códigos, algoritmos y procesos para la extracción de características sería de gran utilidad para conseguir avanzar en este ámbito de estudio. Asimismo, en la mayoría de los estudios, el conjunto de datos a analizar está en inglés y algunos, de manera aislada, en alemán, holandés e italiano. Por lo tanto, es necesario investigar en otros idiomas de uso común en Internet como, por ejemplo, el francés, portugués, mandarín y español (Fortuna & Nunes, 2018).

Sin embargo, en esta investigación vamos a adoptar un enfoque distinto al empleado hasta el momento para el análisis de la comunicación violenta y el discurso de odio, en el que el idioma no será un hándicap. Para ello, utilizaremos un enfoque criminológico, concretamente ambientalista, que nos dará una base teórica para lograr nuestro objetivo: conocer las características de este fenómeno en Internet (concretamente en Twitter) identificando las características ambientales de la comunicación neutral que nos permitan distinguirlas de la comunicación violenta y el discurso de odio mediante la elaboración de un predictivo basado en metadatos.

# PARTE II

# CRIMINOLOGÍA Y DETECCIÓN DE LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO

# Capítulo 3

# LA CRIMINOLOGÍA ANTE LA DETECCIÓN DE LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO EN INTERNET

"People who are criminals in real life will be criminals on the Internet (…) As the Internet moves to the mainstream, all those things will show up. It's just part of the maturation of the medium"

(Bill Gates, 1996)

## 1. ¿Explicar o Detectar? El Papel de la Criminología en la Detección de la Comunicación Violenta y el Discurso de Odio en Internet

Las primeras aproximaciones empíricas al estudio del fenómeno delictivo se centraron en el individuo, en sus motivaciones y en sus condicionantes biológicos y sociales. En definitiva, en las causas de un comportamiento criminal que debían abordarse como si de una enfermedad se tratase. Asumiendo como ciertas estas causas, y con el fin de atajar el mal padecido por el individuo delincuente, los científicos y los poderes públicos se afanaron en diseñar estrategias que debían incidir en el mejor modo de disuadirle por medio del castigo y la intervención del sistema de justicia. También, buscaron la terapia más certera para rehabilitarle y ensayaron el enfoque más adecuado para intervenir en una sociedad desigual generadora de contradicciones, conflictos y

frustraciones. Sin embargo, estos planteamientos, soportados ampliamente por las investigaciones científicas de relevantes autores, fueron cuestionados en los primeros años de la década de los setenta tras constatar un evidente incremento de las tasas de criminalidad. Y es que, tras el desarrollo y aplicación de amplios programas impulsados por los Estados, en los que se invirtieron importantes recursos en la Policía, las prisiones, el sistema de justicia y los servicios sociales, se comenzó a generar un clima de desánimo entre buena parte de los científicos y la sociedad que fue resumido por Ray Jeffery (1977) con un lacónico "todo ha fallado". De este modo, nuevas investigaciones comenzaron a desplazar el foco de atención del individuo al evento delictivo. Importaba, a partir de entonces, el entorno en el que se localizaba el crimen y también, los elementos ambientales que lo favorecían ofreciendo oportunidades que facilitaban el delito. Y así, se iniciaba un paradigma distinto al clásico, de carácter preventivo, que buscaba la anticipación, que consideraba que eran las oportunidades a las que se veían expuestos los individuos las se encontraban en la génesis del crimen. Bajo esta perspectiva, vamos a destinar este capítulo, cuyo fin será contextualizar esta investigación y dotarla de un adecuado marco que permita comprender los razonamientos que se irán viendo.

### *1.1. De la criminalidad al evento criminal*

Las raíces del delito, los comportamientos que quiebran las normas de convivencia que nos hemos dado y, sobre todo, sus negativas consecuencias, han sido y continúan siendo objeto de profundas reflexiones y honda preocupación para las sociedades.

En la etapa precientífica, las explicaciones del delito se buscaban en consideraciones místicas, metafísicas o mágicas. Ya en el hombre primitivo, ciertas conductas que son consideradas como crimen son, a la vez, tabú. Y, por lo tanto, no pueden realizarse. Y el criminal, es decir el que lo ha violado, es automáticamente segregado del grupo social, y

no es raro que él mismo se separe de él. En la antigua Mesopotamia el Código de Hammurabi combatió la corrupción de la Administración, quitó la función judicial a los sacerdotes, y estableció cuestiones como las del cuidado de los delincuentes pobres para que no quedaran desamparados, o el establecimiento de un tribunal superior de apelación. En el antiguo Egipto el Derecho, la religión, la magia y la ciencia eran una misma cosa. Y en el Libro de los Muertos (en la declaración que el muerto debe hacer sobre su vida) puede encontrarse un claro catálogo de todo lo que podía considerarse como antisocial. La antigua China, además de mostrarse como un precedente clásico de la aplicación de la Medicina Legal, impuso la pena proporcional, y su sistema está plagado de detalles que pueden considerarse como humanitarios, en relación con otras civilizaciones asiáticas de la época (Rodríguez-Manzanera, 2007).

Posteriormente, ya en los grandes pensadores de la antigua Grecia, pueden observarse reflexiones sobre los delincuentes y la necesidad de que sean castigados. Y los dramas de la misma época demuestran gran penetración psicológica para los actos criminales que, sin embargo, eran concebidos como una expresión de la acción fatal de los Hados (Seelig, 1958). Rodríguez Manzanera (2007) destaca, entre otros ejemplos más, que Platón ya propugnaba el tratamiento del delincuente y el aspecto preventivo del delito; o que Aristóteles aseguraba que las pasiones pueden ser causa del delito, aún en el hombre virtuoso, y que éste resaltó la influencia criminógena de la pobreza, de la ambición, de los malos hábitos, y de la "razón desviada".

Para el mismo Ernesto Seelig (1958) fue la Medicina Legal la que inicialmente empezó a ofrecer una perspectiva criminológica rudimentaria. Visión que comparte Günther Kaiser (1983), que recuerda que, ya en 1220, el Espejo Sajón declaraba que "*no hay que juzgar al verdadero loco, ni al hombre que carece de juicio*", y que en 1249 se tomó juramento en Bolonia a Hugo de Lucca como médico municipal. Y,

señala que un cambio decisivo se opera en la época del Renacimiento, cuando se estimula la búsqueda del conocimiento a través de la aparición del principio de la investigación de la realidad material, que puede apreciarse en los códigos penales de los siglos XV y XVI.

Una fundamental renovación de las antiguas perspectivas se produce, en el siglo XIX, cuando Rafael Garofalo (1890) fue uno de los primeros en expresar, pública y expresamente, la necesidad de un cambio de paradigma en el estudio de la criminalidad, y la lucha contra ésta, denunciando que "la sociedad no se preocupa del delito tanto como debiera, ni por respecto a la víctima, ni por respecto a la prevención"

Fue a finales del siglo XVIII y principios del XIX cuando, se desarrollaron las primeras investigaciones científicas de la mano de Cesare Beccaria y Jeremy Bentham. El primero, noble italiano milanés, publicó en 1764 su obra "De los delitos y de las penas", en la que criticó la irracionalidad, la arbitrariedad y la crueldad de las leyes penales del siglo XVIII (Garrido & Redondo, 2013). Para él, el concepto de pena era fundamental, ya que gracias a ella se prevenía el delito y de este modo, se protegía el orden social. Por su parte, Jeremy Bentham, filósofo y jurista inglés, en su obra "El Panóptico", publicada en 1780, propuso una cárcel en la que la arquitectura estuviera al servicio del tratamiento, puesto que la finalidad de la pena era la de reformar y corregir al delincuente (Garrido & Redondo, 2013) .

Tal como esquematiza Antonio García-Pablos (2008), junto al pensamiento utópico encabezado por Tomás Moro, y el ilustrado formado, entre otros, por Rousseau, Montesquieu y Voltaire, se unen teóricos sistematizadores de las primeras construcciones científicas, como el anteriormente citado Bentham (en Inglaterra), Filangieri (en Italia), o Feuerbach (en Alemania), para formar el repertorio de ideas sobre el problema criminal que tiene su origen en las llamadas "ciencias del espíritu".

Y, el mencionado autor resume esta corriente de pensamiento señalando que, aun siendo un grupo muy heterogéneo, comparten la imagen del hombre como ser racional, igual y libre; la teoría del pacto social como fundamento de la sociedad civil, y el poder; y la concepción utilitaria del castigo.

Fueron los positivistas, cuyo surgimiento puede situarse en la Italia del siglo XIX, los que dieron la denominación de Escuela Clásica a todas esas corrientes y teorías precedentes a ellos, y como suele suceder, los postulados del nuevo paradigma se enfrentaron violentamente con los del que pretendían sustituir. De este modo, mientras que los autores clásicos situaban al hombre en el centro del universo y, por tanto, comprendían el delito como un acto individual que contravenía a la Ley, los positivistas, entendían que el individuo estaba determinado biológicamente.

Modelos biológicos a los que, según ellos, corresponde explicar científicamente la relevancia criminógena de ciertas variables, ya que determinados datos biológicos diferenciales parecen ser una realidad incuestionable. Y cuestionan las concepciones ambientalistas porque no son capaces de explicar, por sí solas por qué el crimen se distribuye de forma no homogénea, concentrándose en torno a muy reducidos grupos humanos, cuyos individuos acaparan significativamente la mayor parte de los delitos. No obstante, y en este sentido, en la actualidad se asume que es necesario diferenciar entre una perspectiva biológica moderada, de la de un determinismo biológico radical. En el primer caso, sus seguidores sostienen que el sustrato biológico del individuo es un potencial valioso para explicar ciertas conductas antisociales, sin tener que obviar la participación de otros posibles factores; mientras que la opción radical (ya generalmente desestimada) se decanta por la idea de que el delincuente es “distinto”, y porque es distinto, delinque (García-Pablos, 2008).

Interesados en trasladar el método científico desde las ciencias naturales a las ciencias sociales, los positivistas centraron sus análisis en el individuo delincuente y en la etiología del crimen desde perspectivas antropológicas (Bertillon, 1909; Goring, 1913; Hooton, 1931), biotipológicas (Di Tullio, 1980; Kretschmer, 1961; Sheldon, 1942), endocrinológicas (Marañon, 1946; Ruiz-Funes García, 1929), genéticas (Exner, 1957; Jacobs, 1961), o neurobiológicas (Raine, 1993, 2013), etc., con el objetivo final de identificar las características orgánicas y físicas que le hacen diferente al resto de los individuos (García-Guilabert, 2014). Por ejemplo, Lombroso (1876) clasificó a los delincuentes, haciendo referencia, entre otros, al "delincuente nato o atávico", el cual tendría una predisposición delictiva tan fuerte, que la sociedad nada o muy poco podía hacer para evitar que delinquiera. Según Lombroso, las características anatómicas que le diferenciaban del resto eran, la frente huidiza y baja, gran desarrollo de arcadas supraciliares y de pómulos, asimetrías craneales, altura anormal del cráneo, orejas en asa y gran pilosidad (Garrido & Redondo, 2013; Lombroso, 1876; Rodríguez-Manzanera, 2007).

Aunque estas investigaciones hicieron aportaciones muy relevantes, no lograron explicar la variabilidad delictiva, ya que algunos individuos se caracterizaban por tener motivaciones banales o difíciles de comprender. Así surgieron nuevos paradigmas, en respuesta a las limitaciones de los anteriores.

El positivismo psicológico centró su investigación en el psiquismo del delincuente, analizando los posibles trastornos mentales o psicopatologías (Schneider, 1997), su inteligencia (Chico, 1997; Gardner, 1995; Henggeler, 1989), la personalidad (Eysenck, 1989), etc. que pudieran explicar su comportamiento. Los autores se adentraron en la mente del criminal para encontrar una explicación a sus actos. En este sentido, investigadores como el matrimonio Glueck (1956) compararon a delincuentes y no

delincuentes, teniendo en cuenta variables somáticas, capacidad intelectual, temperamento, emociones, etc. Afirmaron que los delincuentes se distinguían por ser menos temerosos, y más asertivos, agresivos y extrovertidos. Por su parte, Eysenck (Eysenck, 1989) formuló la propuesta de las tres dimensiones de la personalidad (neuroticismo, extroversión y psicoticismo), indicando que el delincuente mostraría puntuaciones altas en neuroticismo y extroversión, ya que tendría más problemas en aprender una respuesta que evadiera sus conductas criminales. También postularon (Cleckley, 1941; Hare, 1970) la existencia de un tipo específico de delincuente, manipulador, insensible, mentiroso y sin empatía, que puntuaría alto en la dimensión de psicoticismo, el cual equivaldría al término psicópata, que se relacionaría con delitos más violentos. En cualquier caso, esta orientación de corte psicológico no dejó de lado la relación entre la personalidad criminal y los factores externos, por lo que también analizaron la correlación entre la delincuencia y el entorno familiar (Canter, 1983; Glueck & Glueck, 1950).

Por su parte, el positivismo sociológico comenzó a plantearse la importancia que tenía la sociedad en relación con el delito, contemplando este hecho como un fenómeno social en toda su extensión (García-Guilabert, 2014). A finales del siglo XIX, un grupo de científicos de la Universidad de Chicago (Robert Park, Ernest Burguess, Clifford R. Shaw y Henry D. Mckay) promovió la aplicación del método científico en el estudio del comportamiento humano y social. Estos autores, integrantes de la conocida Escuela de Ecológica de Chicago, abogaron por una sociología más rigurosa y empírica, observando una conexión evidente entre la conducta antisocial y el contexto social donde se realizaba (Garrido & Redondo, 2013). Derivada de sus postulados surgió la teoría de la anomia de Merton (1938), las teorías subculturales delictivas (Cloward & Ohlin, 1960; A. K. Cohen, 1955), las teorías del proceso social (Hirschi, 1969;

Reckless, 1970; Sykes & Matza, 1957) y las teorías de aprendizaje (Akers, 1997; Sutherland, 1947). Esta orientación sociológica, afirma que la delincuencia es fruto de la estructura social, donde existe un desequilibrio entre las metas culturales y las normas institucionales en la sociedad (Durkheim, 1894; Merton, 1938). Debido a esta situación, muchos individuos recurrirían a conductas criminales para lograr objetivos que la sociedad plantea como necesarios (García-Guilabert, 2014). Según su perspectiva, la subcultura delictiva surgirá cuando el grupo de delincuentes tenga comportamientos que para ellos sean dignos de reconocimiento, mientras que, para el grupo de no delincuentes, esos mismos comportamientos deben ser castigados (Cloward & Ohlin, 1960; Cohen, 1955). Esta teoría asume que la criminalidad tenía su origen en concretos grupos juveniles de clase baja. Sin embargo, los defensores de las teorías del proceso social no compartían tales planteamientos, afirmando que cualquier persona, independientemente de la clase social a la que perteneciera, podía convertirse en un criminal (García-Guilabert, 2014). Con las teorías del control surge un nuevo cambio de paradigma en el que la cuestión principal no son las motivaciones que llevan a los individuos a delinquir, sino por qué no lo hacen. Los autores pertenecientes a esta corriente buscaron los mecanismos que llevan a un individuo a no cometer delitos y respetar las normas que rigen una sociedad o, en otros casos, lo contrario (Hirschi, 1969; Reckless, 1970; Sykes & Matza, 1957).

Por su parte, las teorías del aprendizaje social se asentaron en la explicación de la conducta delictiva a partir de una serie de mecanismos de aprendizaje. Desde esta perspectiva, consideraron que era posible modificarla, con la enseñanza de nuevas conductas sociales. En relación con estos enfoques, la teoría de la asociación diferencial de Edwin Sutherland (1947), argumenta que la conducta delictiva se aprende como cualquier otra, independientemente de la clase social del individuo y de su herencia

genética (Garrido & Redondo, 2013). Así, su inclinación delictiva será fruto de un exceso de interacción y comunicación con personas favorables a la infracción de la Ley. En este sentido, Akers (1997), destaca la importancia de la familia, amigos, escuela y otros grupos cercanos como contextos naturales para el aprendizaje. En definitiva, ambos autores, defienden la idea de la influencia que tienen los grupos sociales en el individuo, los modelos a seguir durante su desarrollo, y los refuerzos positivos o negativos que pueda obtener como consecuencia del delito.

Más tarde, a mediados de la década de los 60 del siglo pasado surgen las teorías del etiquetamiento (*labelling aproach*), cambiando radicalmente la criminología positivista que analizaba, fundamentalmente, las causas del comportamiento delictivo. Los autores más representativos de este enfoque, Edwin Lemert, Howar S. Becker y Erving Goffman, se interesaron más en el estudio de la reacción social ante el crimen que en los factores que lo generaban (Garrido & Redondo, 2013). Así, argumentaron que la conducta delictiva dependería de lo que la sociedad entendiera como tal y si por diferentes motivos la persona fue "etiquetada" como un criminal, su autoconcepto se identificaría con el de un criminal. Y, como consecuencia, su forma de actuar se ajustaría a la de un delincuente (García-Guilabert, 2014).

La Criminología Nueva o Crítica, influenciada por los pensamientos de Marx y las teorías del conflicto social, también se enfrentó a las teorías positivistas (Garrido & Redondo, 2013). Taylor, Walton y Young (1973), critican a los teóricos que habían desvinculado al individuo de la sociedad, exigiendo una teoría plenamente social. Esta nueva criminología pretende influir en la política para modificarla, ya que los autores afirmaban que, tanto la desviación social como la delincuencia, estaban relacionadas con el control social, dentro de la lucha de clases y la confrontación entre proletarios y

capitalistas. Tanto esta nueva corriente, como las teorías del etiquetamiento, centrarán su análisis en el acto desviado y, sobre todo, en la reacción social que provoca.

En definitiva, la Escuela Clásica, puso el foco en el estudio de las penas, argumentando que los delincuentes actuaban sobre la base del libre albedrío. Tal como recuerda Serrano Maíllo (2006), hay que reconocer que dicha concepción ha construido el Derecho Penal y la Administración de Justicia contemporáneos (y especiales ejemplos de ellos son los ámbitos español y latinoamericano), al aceptar que la criminología clásica y neoclásica favorecen la investigación sobre los efectos preventivos de las penas. Con respecto a esta última, José Ángel Fernández Cruz (2008) afirma que la denominada Nueva Escuela Clásica (o Escuela Neoclásica) ha rescatado los presupuestos principales de su antecesora: la racionalidad y la búsqueda de la consecución de los propios intereses del criminal, y la teoría de la prevención a través de una pena proporcional, cierta y rápida.

Por su parte, los positivistas, pusieron el acento en las perspectivas biológicas, psicológicas y sociológicas, tratando de explicar de este modo el actuar desviado. Y también se puede comprobar, en la actualidad, una evidente reformulación y renacimiento de esta corriente de pensamiento criminológico. En palabras de Akers, "*las explicaciones biológicas del delito han llegado a ocupar un nuevo lugar de respeto en criminología*" y "*se toman más en serio hoy, que en cualquier otro momento desde la primera parte del siglo XX*". Y, efectivamente, se puede comprobar un importante cuerpo de evidencia empírica sobre los factores genéticos y biológicos que intervienen en la criminalidad (Serrano-Maíllo, 2006). Y, en este aspecto, hay que destacar el creciente protagonismo de las modernas investigaciones sobre Neurocriminología, que se han añadido decididamente a las más tradicionales, relacionadas con las

antropológicas, antropométricas, biotipológicas, endocrinológicas, genéticas, o bioquímicas.

En todo caso, la evolución de los diferentes enfoques no ha modificado su objeto de estudio, centrado siempre en el individuo. Y hay que reconocer que un enfoque excesivamente reduccionista de esta perspectiva limita gravemente la posibilidad de la existencia de una concepción deseablemente holística del concepto del delito y la criminalidad. Es por ello por lo que, tal como afirma Prieto (2011), el objeto de estudio de la criminología, y la complejidad del mismo, promovió la incorporación de nuevos elementos de estudio, como la víctima o el ambiente.

### *1.2. La criminología ambiental y el evento criminal*

De este modo, en la década de los 60 y fundamentalmente en la de los 70 del siglo pasado, se produjo un cambio de paradigma, dirigiendo esta vez el foco de atención al estudio del evento criminal (Jacobs, 1961; Jeffery, 1977; Newman, 1972). La obra *The death and life of great american cities* publicada en 1961 por la urbanista y escritora Jane Jacobs, planteó la importancia que tiene confección del entorno urbano en la génesis del delito. Si es adecuado, favorecerá la presencia de las personas en las vías públicas, ejerciendo una vigilancia natural ante los posibles delitos, lo que ella denominó "ojos en la calle" (Medina-Sarmiento, 2013). Por su parte, el arquitecto Oscar Newman, propuso el concepto "espacio defendible", en su publicación de 1972 *Defensible Space: Crime Prevention Through Urban Design*, haciendo referencia a la adecuada configuración de ambientes seguros (zonas iluminadas, control de accesos, etc.) como estrategia de prevención ante el crimen.

Pero fue Ray Jeffery el autor más crítico con las políticas preventivas del delito, y con los fracasos del sistema de justicia desarrollados en la época de la criminología

tradicional, donde el énfasis se encontraba en la rehabilitación del delincuente y el tratamiento (Medina-Sarmiento, 2013). En su obra, *Crime Prevention Through Environmetal Design* (1977), propone un nuevo modelo preventivo, donde lo importante es reducir el delito mediante medidas implantadas antes de que éste se produzca, lo que debe entenderse como prevención primaria, basando su argumento en el adecuado diseño de espacios (Jeffery, 1977). En esta misma línea ambientalista, surge la teoría de la elección racional del delito, formulada por Dereck Cornish y Ronald Clarke (1986), en la que critican la visión tradicional del delincuente, considerando que el comportamiento delictivo no es fruto de la enfermedad mental, ni de problemas biológicos o hereditarios que se mantienen estables en el tiempo (Medina-Sarmiento, 2013). En este caso, se deja atrás el pensamiento convencional y se introducen elementos ambientales en la toma de decisiones del delincuente (Cornish & Clarke, 2008). Los autores afirman que el delito va a depender, en cada momento concreto, de la valoración que haga el individuo de su entorno, y entre los costes de su acción (detención) y los posibles beneficios (dinero, sexo, etc.). El individuo debe estar dispuesto a cometer el delito, y tendrá que decidir, en función de los factores situacionales, si cometerá un delito aislado o iniciará una "carrera delictiva" (Medina-Sarmiento, 2013; Medina, 2011). La valoración que realiza el delincuente en todo el proceso, aunque se suele denominar "racional", realmente no lo es del todo, ya que normalmente será el resultado de un análisis subjetivo de todas las variables que le rodean.

Estos elementos son importantes porque, si tradicionalmente las medidas de carácter preventivo iban enfocadas al tratamiento del criminal, gracias a esta corriente se formulan nuevas estrategias preventivas, que inciden más en el ambiente. Por ejemplo, en la creación de obstáculos (sistemas de alarma, cámaras de seguridad, etc.) para que,

en el análisis que realice el delincuente, los costes excedan a los beneficios. En definitiva, la teoría de la elección racional ofrece una forma de analizar la delincuencia, más centrada en el presente y en la influencia del ambiente en este comportamiento.

La principal debilidad de la mayoría de las teorías criminológicas ha sido equiparar la criminalidad con el crimen, cuando la criminalidad es sólo uno de los elementos que constituye el evento delictivo (Brantingham & Brantingham, 1993). De este modo, la corriente ambientalista unió las concepciones de la teoría de Cornish y Clarke (Cornish & Clarke, 1986) y la teoría del patrón delictivo (Brantingham & Brantingham, 1991).

En esta última teoría, los autores abandonan la concepción generalizada de la aleatoriedad del crimen en el espacio y en el tiempo, incluso en la sociedad, argumentando que los crímenes están relacionados entre sí, y por lo tanto existen patrones definidos para cada tipo de delito. Lo que dependerá de los grupos sociales o vecindarios, así como de la vida rutinaria de un individuo (Brantingham & Brantingham, 1993). Analizan el crimen como un fenómeno complejo, como un evento que puede ser consecuencia de diversas causas. Y así existirán zonas donde se concentren gran cantidad de delitos (puntos calientes o *hot spots*), y otros, en los que apenas haya delincuencia (puntos fríos o *cold spots*). Además, existirán delincuentes y víctimas reincidentes (Brantingham & Brantingham, 2013).

En definitiva, para ellos la delincuencia está agrupada en función de las zonas donde viven los individuos, cómo y por qué viajan por la ciudad, y si es por trabajo, ocio, etc. En este sentido, habrá concentraciones de actividades (mercados, centros comerciales, zonas de bares, etc.), que pueden ser generadoras del crimen, y otras que puedan atraerlo. Los patrones son dinámicos, pero teniendo en cuenta los elementos analizados, se pueden comprender las pautas de la criminalidad, y de este modo diseñar medidas preventivas que puedan reducirla.

Tanto la teoría de la elección racional, como la teoría del patrón delictivo, se inscriben dentro de las denominadas "teorías del crimen" (Gottfredson & Hirschi, 1990). Dentro de este grupo de teorías, se debe incluir el estudio de la Teoría de las Actividades Cotidianas de Cohen y Felson (1979), paradigma de las teorías de la oportunidad. Junto a esta teoría, se unirían las del patrón delictivo (Brantingham & Brantingham, 1981) y la de la elección racional (Cornish & Clarke, 1986), creándose las teorías de la oportunidad. Este nuevo enfoque, como ya se ha mencionado, se centra en el análisis del evento delictivo en su conjunto, en lugar de darle protagonismo único a las características del delincuente, como sucedió en las teorías desarrolladas de la criminología tradicional. Así, esta nueva forma de analizar el delito se ha ido extendiendo cada vez más. En este sentido se debe indicar que España, comparada con otros países anglosajones, ha tardado más en aplicar el enfoque ambiental en sus estudios científicos sobre el crimen (Vozmediano & San Juan, 2010).

### 1.2.1 *La Teoría de las Actividades Cotidianas*

Según Serrano-Maíllo (2006), las grandes teorías son excesivamente amplias y abstractas, demasiado especulativas y difíciles de entender, desconectadas de la observación. Frente a éstas, las teorías de alcance medio deben ser lo menos abstractas posible, y lo más conectadas que se pueda con la realidad. Así, estas últimas se decantan por formulaciones más próximas a la observación y que, por lo tanto, se mostrarían más útiles.

Varios autores, entre otros, Garcia-Pablos (2008), Fernández-Cruz (2008) y Serrano-Maíllo (2017) incluyen dentro de la concepción de la Escuela Liberal Clásica (que no se preocupa especialmente de la etiología del crimen, o de la consideración del delincuente como un ser diferente, sino sobre el hecho entendido como una violación del pacto social), varios submodelos. Los más importantes, el de las teorías del entorno físico, el

de la teoría de la elección racional (o teoría económica), y la visión de la teoría de las actividades cotidianas. En concreto, Fernández-Cruz (2008) sostiene que las dos últimas se complementan recíprocamente: la segunda podría definirse como una macro-teoría (gran teoría en terminología de Serrano Maíllo) que trata de poner de manifiesto las relaciones entre el crimen y las actividades cotidianas en la sociedad, mientras que la anterior es una micro-teoría (teoría de alcance medio, volviendo a la terminología de Serrano Maíllo) que detecta y evalúa cómo estas oportunidades son percibidas por los criminales para determinadas categorías de delitos. Y, en función de ello, se deriva la necesidad de profundizar, a continuación, en ambas perspectivas de forma complementaria.

A pesar de la mejora en las características sociales y económicas en los EE. UU. después de la II Guerra Mundial, la tasa general de delincuencia aumentó de manera aparentemente inexplicable. Por ello, Lawrence Cohen y Marcus Felson se plantearon analizar este fenómeno desde una perspectiva distinta hasta el momento. Así, en 1979, en su publicación "*Social Change and Crimen Rate Trends: A Routine Activity Approach"*, examinaron la tendencia de la tasa de delincuencia en los EE. UU., entre los años 1947 y 1974, observando que ciertamente había aumentado la criminalidad a pesar del bienestar social que existía. Argumentaron que los cambios estructurales en los patrones de las actividades cotidianas de las personas en las ciudades, la dispersión de las actividades lejos de los hogares y las familias, así como los cambios tecnológicos en bienes y servicios, podrían haber influido en ese aumento de la delincuencia. Sobre todo, al coincidir la confluencia en el espacio y en el tiempo de tres elementos fundamentales: un delincuente motivado, un objetivo o víctima adecuados, y la ausencia de un guardián capaz de evitar el delito. Además, inciden en la importancia de la

concurrencia de estos tres elementos porque, según los autores, la falta de cualquiera de ellos es suficiente para evitar la comisión de un crimen (Cohen & Felson, 1979).

#### 1.2.1.1 ¿Delincuentes motivados?

El delincuente motivado, como lo denominaron Cohen y Felson (1979) en su primera publicación, hace referencia a cualquier persona que pueda tener un motivo para delinquir y cualidades para ello (Felson & Cohen, 1980). No niegan la importancia de los factores que motivan a los criminales para cometer los delitos, pero centran su análisis en el evento delictivo y, sobre todo, en las circunstancias previas al mismo (Cohen & Felson, 1979). Teniendo en cuenta cuál era el interés de estudio para los autores, en las siguientes publicaciones abandonaron el término "delincuente motivado" y lo reemplazaron por "delincuente potencial o probable" (Felson & Boba, 2010; Felson & Cohen, 1980; Felson & Clarke, 1998). En este sentido, cualquier persona puede ser un delincuente, no sólo por su motivación o disposición al crimen, sino también porque sus factores físicos, como la fuerza o la agilidad, pueden hacer que se involucre más fácilmente en estas conductas (Clarke & Felson, 2011). Del mismo modo, confluye en este punto la teoría de la elección racional de Cornish y Clarke (2008), donde afirman que el comportamiento delictivo no se puede considerar como el resultado de motivaciones delictivas estables. Sino que los deseos, las preferencias y los motivos de los delincuentes, (y posibles delincuentes), son similares a los del resto de personas que no delinquen, estando en constante interacción con las oportunidades del entorno para crear, reforzar o incluso reducir los actos criminales. En cualquier caso, ello dependerá de cómo el delincuente potencial interactúe con el resto de elementos formulados en el enfoque de las actividades cotidianas.

#### 1.2.1.2 El guardián en la teoría de Felson

El tercer y último elemento desarrollado en la teoría de Cohen y Felson es la ausencia del guardián capaz: alguien que puede influir para impedir o detener un hecho criminal (Felson & Cohen, 1980). Es decir, sería aquella persona que con su presencia podría evitar el delito o que, con su ausencia, podría aumentar las probabilidades de su comisión (Felson, 1995). Este atributo haría referencia a cualquier individuo que camine por una zona o incluso cuya función sea la de vigilar a las personas o cualquier propiedad (Felson, 2006). Respecto de él, no se debe pensar únicamente en policías y personal de seguridad, sino también en personal que, aunque no esté formado en estos ámbitos, puedan disuadir al delincuente porque podría identificarlo o frustrar su acción. Es cierto que los miembros y fuerzas de seguridad, así como los vigilantes de seguridad, ejercen de manera ejemplar la función de guardián capaz, pero su ausencia es lo más común en el momento de suceder el crimen (Felson & Boba, 2010; Felson, 1986). Sin embargo, no es posible que se encuentre un miembro de estas características en cada rincón de la ciudad para evitar la comisión de delitos. Por ello, se consideran de gran importancia en la prevención, como ya se ha comentado, otros elementos no tan específicos, como pueden ser los dueños de una residencia, un familiar, o amigo, o un viandante, en el desarrollo de sus actividades cotidianas. Los cuales puedan, con su presencia, defenderse a sí mismo y a otros, así como a los bienes o propiedades que estuvieran en peligro.

La concepción de este último elemento, la ausencia de guardián capaz ha sido objeto de debate desde que se formuló. A lo largo de estos años el concepto ha ido evolucionando (Hollis et al., 2013) y aunque en un principio únicamente se tenía presente al individuo como guardián capaz, se han introducido otros elementos que

pueden ejercer la misma función, como podrían ser, por ejemplo, los circuitos cerrados de televisión (CCTV) (Hollis et al., 2011).

Felson (1986) reformuló el concepto de guardián capaz, vinculando el enfoque de las actividades cotidianas con la teoría del control de Hirschi (1969), estableciendo dos pasos:

El primero, hace referencia a las vinculaciones sociales que existen en cada individuo, como por ejemplo la desarrollada entre padres e hijos. Los padres se convierten en "controladores" del comportamiento del niño.

El segundo paso sería la identificación exacta de quién está incumpliendo o va a incumplir las leyes. Esto puede ser sencillo en pueblos pequeños, pero en una gran urbe se complica más, y los responsables serían aquellos actores sociales encargados de vigilar el espacio, como conserjes, comerciantes, taxistas, etc. (Eck, 1994; Felson, 1995).

Mientras que el triángulo original del evento delictivo desarrollado por Cohen y Felson (1979) contiene tres elementos fundamentales (delincuente motivado, objetivo adecuado y ausencia de guardián capaz) que confluyen en el espacio y en el tiempo, la evolución del concepto de guardián capaz desde la perspectiva del control social informal (1995), junto con los planteamientos de John Eck (1994), derivó en una revisión de este esquema planteándose una versión ampliada del conocido triángulo del crimen (Figura 4).


*Figura 4. Triángulo del crimen. Adaptado de John Eck, 1994. Center for Problem Oriented Policing (2013)*

La reformulación presenta inicialmente los tres lados del triángulo: objetivo o víctima del crimen, el delincuente y el lugar donde ocurre el evento. Lo que muestra es que el delincuente y el objetivo del delito se encuentran en un lugar adecuado que permite la comisión del evento delictivo (Hollis et al., 2013). Luego, por fuera de este triángulo, habría otro en el que los responsables de cada uno de estos elementos podrían reducir la probabilidad de que ocurriera el evento criminal realizando acciones preventivas. De este modo, el cuidador vigila al delincuente, los encargados del lugar (responsables o gestores) controlan el entorno y los guardianes o vigilantes supervisan los objetivos (Hollis et al., 2013).

En cualquier caso, para que el delincuente tenga éxito en su acción criminal deberá liberarse de sus cuidadores y encontrar un objetivo desprotegido por los guardianes, en un lugar que no esté vigilado por los gestores (Felson, 1995). Pero para que estos nuevos elementos considerados tengan éxito en la prevención del crimen, Felson (1995) indica que deberán relacionarse con su grado de responsabilidad, estableciendo cuatro niveles: el personal (amigos, propietarios y familia), el asignado (empleados, vigilantes, etc.), el difuso (trabajadores con responsabilidades muy generales) y el general (cualquier individuo).

En este sentido, Felson (1998) incide en la idea de que los delincuentes cometen delitos con el mínimo esfuerzo posible. Por lo tanto, si estos elementos consiguen que el crimen sea más difícil de cometer, el delincuente será menos propenso a delinquir (Levy, 2009).

Sobre la base de esta concepción, Sampson, Eck y Dunham (2010) investigan las razones que pueden llevar a los controladores a no ser efectivos. En este sentido, desarrollan el concepto de los "supercontroladores", haciendo referencia a las personas, organizaciones e instituciones que generan incentivos para que los controladores influyan en la prevención del delito o por el contrario lo faciliten. Los "supercontroladores" solo tienen efectos indirectos sobre las condiciones necesarias para la delincuencia. De este modo, se incluiría un tercer triángulo al esquema propuesto por Felson (1995) y Eck (1994), con diferentes tipologías de "supercontroladores" agrupadas en tres categorías (Sampson et al., 2010):

1. Formal: Organizativo, contractual, financiero, normativo, tribunales de justicia.
2. Difusa: Política, mercados y comunicación
3. Personal: Grupos y familia

Estos "supercontroladores" dependen de la autoridad social, legal o financiera formal para modificar el comportamiento de los controladores. Además, ejercen esta autoridad dentro de un marco institucional establecido que define quién influye en quién, de qué manera y bajo qué circunstancias (Sampson et al., 2010).

Aunque el concepto de "súpercontrolador" esté vinculado a una mejor prevención del crimen, no debe confundirse con que ya se haya conseguido el éxito en ella ya que, incluso si los "súpercontroladores son seleccionados con éxito, los esfuerzos y las estrategias de los controladores puede que no sean apropiadas. En definitiva, los

"supercontroladores" pueden ser una condición necesaria para iniciar acciones preventivas, pero esto no significa que sean suficientes para una prevención eficaz, a no ser que los controladores utilicen métodos adecuados para prevenir o reducir el crimen.

#### 1.2.1.3 Los elementos ambientales del crimen

El objetivo adecuado hace referencia tanto a una persona como una propiedad que pueden ser atacadas por un criminal (García-Guilabert, 2014). La decisión de emplear el concepto "objetivo" en lugar de "víctima" fue una decisión tomada por Cohen y Felson, para dejar constancia de que, aunque la victimología era muy conocida en ese momento, el concepto "víctima" no diferenciaba a las víctimas de las agresiones directas de aquellas víctimas indirectas de otro tipo de delitos, como robos o hurtos en los que la víctima no estaba presente (Felson & Clarke, 1998). Como ya se ha indicado, el enfoque de las actividades cotidianas se centra en el punto de vista del agresor, no en el de la víctima y tampoco en el de la sociedad. Por tanto, desde la perspectiva del criminal que comete delitos contra la propiedad, el objetivo del delito será un bien y no una persona concreta (Felson & Cohen, 2008). En cualquier caso, el delincuente potencial tendrá que percibir estos objetivos como adecuados. Y para ello, valorará si la posición en el espacio y en el tiempo es favorable además de otros cuatro elementos fundamentales que determinan su nivel de riesgo, asociados al acrónimo VIVA: *Value, Inertia, Visibility* and *Access* (Cohen & Felson, 1979; Felson & Clarke, 1998).

- *Value* (Valor): Este elemento hace referencia tanto al valor real del objetivo, como al simbólico que puede tener para el delincuente.
- *Inertia* (Inercia): El delincuente valorará si es adecuado en función de su forma, tamaño o peso. Si el objetivo es una persona, tendrá en cuenta que pueda

neutralizarla sin perder en el enfrentamiento. Si es un bien, lo importante será que pueda transportarlo sin riesgo alguno.

- *Visibility* (Visibilidad): Este atributo se refiere a la exhibición de los objetivos ante el delincuente, como por ejemplo dejar unas gafas de sol dentro de un coche a la vista de todo el mundo.

- *Access* (Accesibilidad): Esta característica se basa en el diseño del entorno (calles, parques, comercios, etc.) y la ubicación del objetivo, aumentando o disminuyendo el riesgo para el delincuente.

Cohen y Felson (1979) describieron VIVA de manera muy sucinta, centrándose en la idea del crimen como evento delictivo ocurrido en el espacio físico. Y por ello, debía explicarse desde una perspectiva ecológica (Clarke, 1999). Para distanciarse todavía más de la criminología tradicional, no realizan diferencias entre víctima de delitos y objetivos inanimados, ya que este acrónimo, según los autores, se puede aplicar tanto a las víctimas de agresiones como a los objetivos del robo. A lo largo de los años, el enfoque de las actividades cotidianas se ha ido desarrollando y complementando con otras teorías, como la formulada por Cornish y Clarke (1986) sobre la elección racional del delincuente. En este sentido, Clarke (1999) argumenta que VIVA tiene muchas limitaciones, al pretender cubrir todos los objetivos de los delitos y no valorar la motivación del delincuente. De este modo, para determinados tipos de delitos ofrece una modificación de VIVA, dando como resultado el modelo CRAVED (por sus siglas en inglés) de objetivos de robo que introduce nuevos términos como el de *hot products*, para hacer referencia a los objetos más atractivos para los ladrones.

El acrónimo CRAVED (*conceable, removable, available, valuable, enjoyable and disposable*) describe las características de los productos que se roban con más frecuencia (Clarke, 1999).

- *Concealable* (Ocultable): Este elemento hace referencia a los objetos que, una vez robados, son fáciles de esconder y pasan desapercibidos, disminuyendo el riesgo de identificación de los ladrones.
- *Removable* (Extraíble): Al igual que reconoce VIVA, los productos que se transportan más fácilmente tienen más posibilidades de ser objeto de robo.
- *Available* (Disponible): Este atributo recoge dos elementos de VIVA, visibilidad y accesibilidad. Es una condición necesaria para considerar un objetivo como *hot product*.
- *Valuable* (Valioso): Del mismo que VIVA, el valor del objeto es importante. Los ladrones se fijarán más en los productos valiosos, sobre todo si tienen intención de venderlos una vez los hayan sustraído.
- *Enjoyable* (Disfrutable): Este elemento hace referencia a la capacidad de entretenimiento que tendrá para el ladrón. Los delincuentes que roban en barrios residenciales son más propensos a sustraer televisores y vídeos en vez de hornos o microondas, aunque sean igual de valiosos o estén disponibles lo mismo (Clarke, 1999). En general, los "productos calientes" según la investigación, son agradables de poseer y/o consumir, como el tabaco y el alcohol.
- *Disposable* (Disponible): Aunque parezca obvio, muchos artículos se roban para venderlos a otras personas, por lo tanto, los criminales buscarán esta cualidad en los productos.

VIVA fue por tanto un intento de resumir las características de la mayoría de los objetivos seleccionados por los delincuentes en general, evitando los elementos más subjetivos de la elección de los mismos. Además, se formuló antes de la publicación de muchas investigaciones sobre los *hot products*. Por lo tanto, la evolución de VIVA a CRAVED ha sido fruto de un desarrollo para adecuar el modelo a los objetivos, en este caso, de los robos.

Como ha quedado de manifiesto en este apartado, el enfoque de las actividades cotidianas para el análisis del crimen ha evolucionado y se ha complementado con otros enfoques de análisis del delito, explicando gran parte de la delincuencia común en el entorno que nos rodea. En el siguiente punto analizaremos si puede explicar también la delincuencia en el ciberespacio.

## 2. La Oportunidad Delictiva en el Ciberespacio

### *2.1 El ciberespacio como un nuevo ámbito de oportunidad delictiva*

La teoría de las actividades cotidianas analiza, desde un enfoque ecológico, el delito. La capacidad de identificar a los delincuentes y a los objetivos en el espacio y en el tiempo es fundamental porque, como ya se ha indicado en el capítulo anterior, el crimen se produce cuando el delincuente potencial y el objetivo adecuado convergen en el espacio y en el tiempo, en ausencia de un guardián capaz (Leukfeldt & Yar, 2016). Sin embargo, en el entorno virtual la conceptualización del evento delictivo se configura de forma muy diferente, ya que es "antiespacial" (Mitchell, 1995). Es decir, hay distancia cero entre dos puntos cualquiera en el ciberespacio. Y, por lo tanto, es complicado hablar de convergencia entre delincuentes y objetivos en este ámbito (Leukfeldt & Yar, 2016). De este modo, es importante hacer un análisis de las características espacio y

tiempo en el ciberespacio, para conseguir entender la confluencia de los delincuentes y los objetivos en este nuevo entorno virtual.

### 2.1.1 *Las características intrínsecas del ciberespacio: el tiempo y el espacio*

Originalmente, la teoría de las actividades cotidianas explicaba las oportunidades delictivas cuando los criminales convergen con objetivos adecuados en el mismo lugar y con la ausencia de un guardián capaz. Actualmente, sin embargo, Internet y los dispositivos móviles crean nuevas oportunidades delictivas en diferentes condiciones (Brady, Randa, & Reyns, 2016). Por su parte, Yar (2005) argumenta que los delitos cometidos mediante el uso de la tecnología se cometen a distancia y, en general, sin ninguna interacción física directa entre el delincuente y el objetivo. Así pues, existe una divergencia de tiempo y espacio entre el delincuente y el objetivo. Yar sostuvo que la teoría de las actividades cotidianas es limitada en su capacidad de explicar la cibervictimización, ya que la propia divergencia niega los componentes teóricos de la teoría (Brady et al., 2016). La aparición de las TIC e Internet ha planteado interrogantes sobre la adaptación de las teorías criminológicas tradicionales para explicar el crimen y la delincuencia en el mundo virtual.

Reyns y sus colegas (2011) señalaron que el tiempo y el espacio pueden no ser tan relevantes como otros factores que proporcionan oportunidades delictivas. Como, por ejemplo, una mayor proximidad y exposición a delincuentes potenciales y al objetivo adecuado, así como mayor participación *online*. Además, los avances en las nuevas tecnologías sugieren que el tiempo es una constante y es, a través de estas conexiones, donde las vulnerabilidades y oportunidades delictivas aumentan, debido a que los objetivos están muy cerca y expuestos continuamente a delincuentes potenciales que existen dentro de las mismas redes (Brady et al., 2016).

El concepto de convergencia en el espacio y en el tiempo cuando se discute sobre la teoría de las actividades cotidianas, ha cambiado. Con la aparición de Internet y de las TIC, ahora los usuarios están constantemente conectados a una red, lo que da lugar a una exposición continua a posibles delincuentes (Brady et al., 2016). Además, la teoría de las actividades cotidianas adaptada al ciberespacio, sostiene que la separación de los delincuentes potenciales y los objetivos adecuados en el tiempo puede conciliarse considerando su interacción como "rezagada" en el tiempo (Reyns, 2017).

Contrariamente a los delitos predatorios de contacto directo previstos originalmente por Cohen y Felson (1979), muchos ciberdelitos no requieren un contacto inmediato entre la víctima y el delincuente (Reyns, 2017). Por ejemplo, un acosador puede enviar a una víctima potencial un mensaje a través de un correo electrónico, y que ésta no lo lea hasta un tiempo después. Por lo tanto, no hay interacción en tiempo real entre las partes. Sin embargo, esto no excluye la posibilidad de que el receptor se convierta finalmente en víctima de un delito de acoso. Reyns y sus colegas (2011) explicaron que cuando las acciones del infractor en el tiempo 1 son recibidas por el objetivo en el tiempo 2, se produce la interacción "temporal" requerida (Figura 5). En general, el solapamiento temporal entre los delincuentes y las víctimas puede ser inmediato o se puede retrasar en el tiempo (Reyns, 2017).

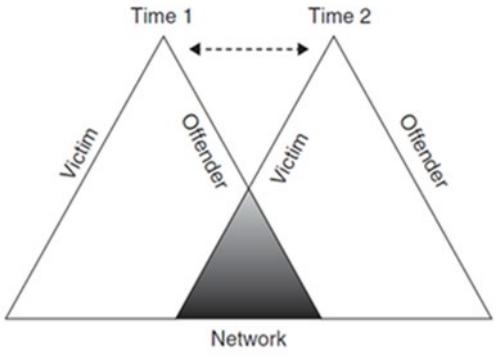


*Figura 5. Intersección asincrónica de las víctimas y los delincuentes en los entornos online (Reyns, 2010)*

En este mismo sentido, Miró-Llinares (2011) y Yar (2005) comparten el cambio de las características espacio-temporales en el ciberespacio, modificando las condiciones del evento delictivo. Sin embargo, Miró-Llinares (2011) difiere de Yar (2005), al argumentar que el enfoque de las actividades cotidianas es idóneo para desarrollar el fenómeno de la cibercriminalidad. La teoría de las actividades cotidianas resalta el lugar como característica fundamental del delito y no el delincuente, lo que posibilita su adaptación. Este nuevo ámbito de oportunidad criminal tiene unas características intrínsecas y extrínsecas diferentes al espacio físico, que debemos conocer en profundidad para comprender de manera adecuada el evento criminal en el ciberespacio.

Así, Miró-Llinares (2011) argumenta que el ciberespacio es un espacio, porque en él los usuarios se reúnen e interaccionan, pero "... mientras que el espacio físico existe antes y seguirá existiendo después de que termine la relación, el ciberespacio agota su existencia en cuanto el mismo sirva para la comunicación entre los sujetos, dado que sin interacción no hay red" (p.6).

También señala Miró-Llinares (2011) que:

> "El ciberespacio es real en el sentido de que existe, pero se trata de una especie nueva de espacio, invisible a nuestros directos sentidos y en el que las coordenadas espacio-tiempo adquieren otro significado y ven redefinidos su alcance y límites. El ciberespacio supone una contracción total del espacio (de las distancias) y, a la vez, la dilatación de las posibilidades de encuentro y comunicación entre personas." (p.6)

Internet ha eliminado las distancias, contrayendo el mundo y aproximando a un mismo lugar en el ciberespacio, a usuarios que pueden estar separados físicamente por miles de kilómetros (Figura 6).

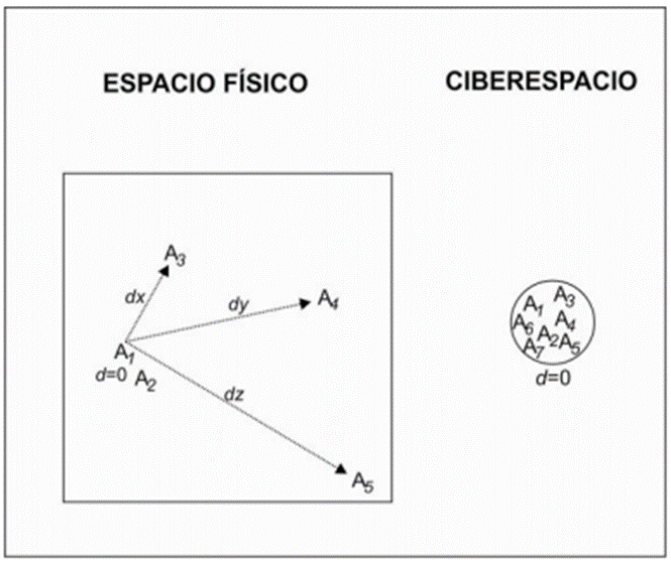


*Figura 6. Contracción de la distancia en el ciberespacio y expansión de la capacidad comunicativa (Miró-Llinares, 2011)*

Internet ha creado un nuevo espacio virtual en el que el concepto de tiempo también ha cambiado. La consecuencia de que el espacio se contraiga es un aumento de la significación del tiempo. De este modo, como argumenta Miró-Llinares (2012), determinadas acciones que para desarrollarse pueden llevar mucho tiempo en el espacio físico, en el ciberespacio se pueden realizar de manera inminente, ya que en el mundo virtual los eventos acontecen más rápido que en espacio físico (Figura 7). En palabras de Miró-Llinares (2012) "...al no requerirse en el ciberespacio recorrer una distancia para la comunicación, las posibilidades de contacto con múltiples sujetos aumentan y se reduce el tiempo necesario para ello." (p.149).

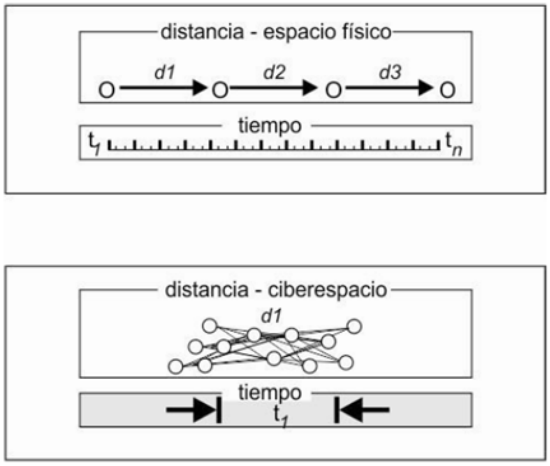


*Figura 7. Contracción del tiempo (Miró-Llinares, 2011)*

Además de la inmediatez en la comunicación de las personas a través del ciberespacio, se debe de tener en cuenta la existencia de los efectos perennes. Es decir, en el espacio físico las comunicaciones se producen de forma momentánea y efímera. Sin embargo, en el mundo virtual pueden permanecer en el tiempo de manera indeterminada, como por ejemplo las publicaciones de contenidos en redes sociales, blogs, etc. En estos casos, los efectos podrían afectar al usuario en el mismo momento de la realización, pero también en el futuro (Miró-Llinares, 2012;2011).

En definitiva, debemos ser conscientes que en el mundo virtual las dimensiones espacio-temporales se modifican de una manera muy significativa, ya que las distancias se comprimen y el tiempo que supone recorrerlas disminuye (F. Miró-Llinares, 2012). Asimismo, si comparamos el espacio físico con el ciberespacio, éste es más vasto y también más permanente.

### 2.1.2 *Las características extrínsecas del ciberespacio*

La configuración de las características intrínsecas del ciberespacio (el tiempo y el espacio) es diferente al espacio físico. Junto a estos caracteres intrínsecos, el

ciberespacio está configurado por otros extrínsecos (Miró Llinares, 2011). Su transnacionalidad, neutralidad, descentralización, deslocalización, universalidad y anonimato, son singularidades que han potenciado su popularización. Pero también han impuesto límites a la persecución y prevención de los crímenes que él acontecen.

2.1.2.1 Transnacionalidad, neutralidad, descentralización y deslocalización

La transnacionalidad, como argumenta Gómez-Águilar (2005), puede que sea la característica más significativa del ciberespacio, concretándose en la ausencia de distancias o fronteras que lo delimiten. De este modo, como indica Miró-Llinares (2012; 2011) cualquier usuario de un Estado nacional puede acceder a cualquier servidor en otro Estado nacional. Por lo tanto, los contenidos de las diferentes páginas web pueden ser vistos por cientos de usuarios en diferentes lugares del mundo. Este hecho supone una propagación de la comunicación nunca vista hasta ahora. Teniendo en cuenta estas características, la transnacionalidad del ciberespacio obstaculiza la investigación de la cibercriminalidad.

El ciberespacio también se caracteriza por la neutralidad, ya que permite a cualquier individuo el acceso a su contenido sin censuras, y su comunicación desde un punto a otro del ciberespacio sin modificar el contenido (Alcantara, 2011; Suárez, 2013). Relacionadas con estas dos características (transnacionalidad y neutralidad), podemos enumerar también como características extrínsecas, pero que configuran el ciberespacio, la descentralización y la deslocalización. Las cuales hacen referencia a la incapacidad de situar Internet en un espacio físico concreto, extendiéndose por todos los lugares a la vez. En este sentido, el concepto de autoridad centralizada en Internet es inexistente. Ni tan siquiera es posible, por parte de instituciones u órganos de control del contenido que circula por el ciberespacio, implantar alguna forma de vigilar la información (Romeo-Casabona, 2006). Teniendo en cuenta que Internet no se rige por normas nacionales de

un único Estado, ni siquiera por unas leyes reconocidas por todos los que lo configuran, los controles por parte de los gobiernos no serán efectivos, al existir diferentes maneras de evitarlos (Miró-Llinares, 2011).

De este modo, como señala Miró-Llinares (2011) "... la existencia de este espacio transnacional, neutro y distribuido, con las consecuencias que conlleva, produce una tensión, en este caso en el plano jurídico, con la casi contradictoria existencia de Estados nacionales con legislaciones distintas reguladoras de este u otro fenómeno." (p. 12).

2.1.2.2 El ciberespacio universal y popularizado

Internet y las Tecnologías de la Información y Comunicación se han convertido en el método de interacción personal más popular en la actualidad, así como en un instrumento esencial no solo a nivel profesional sino también personal. De hecho, cada año aumenta el número de usuarios de Internet en el mundo. Según datos presentados por "We are social" junto con "Hootsuite"[5] en 2015 la cifra total era de 2.831 millones de usuarios en el mundo. En el año 2016 aumentó un 11%, llegando a los 3.153 millones. A principios de 2020 se alcanzaron los 4.538 millones, con una subida del 7% respecto del año anterior con 4.241 millones de usuarios, siendo el 57% de la población mundial la que utiliza internet (Figura 8).

[5] https://wearesocial.com/es/digital-2020-espana

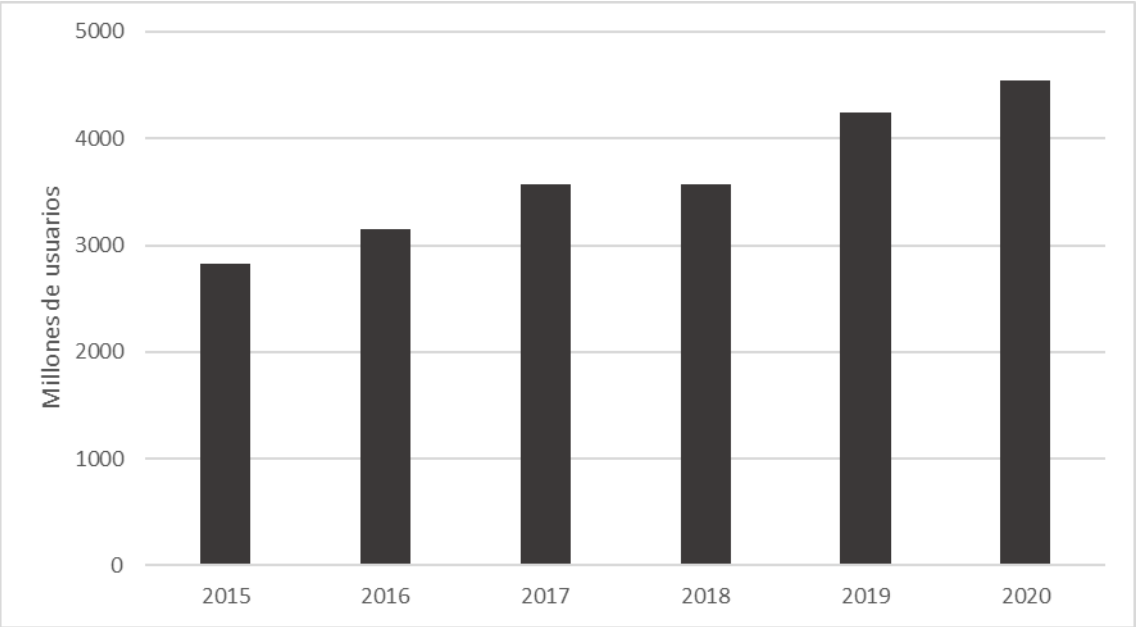


*Figura 8. Nº de usuarios de Internet, en el mundo, por año (en millones) con variación interanual (Elaboración propia a partir de datos de "we are social" y "hootsuite", 2020)*

Además, en estos estudios realizados cada año por "We are social", se enfatiza que, a medida que pasan los años, los usuarios de Internet prefieren navegar desde sus móviles, ya que el 92% de los casos lo hacen a través de este medio. Los usuarios navegan por la Red más de 6 horas al día, lo que supondría algo más de 100 días al año. En España, somos casi 43 millones de usuarios los que empleamos diferentes dispositivos para acceder a Internet y estamos conectados una media de 5 horas y 41 minutos al día (we are social, 2020).

Teniendo en cuenta estas cifras, es innegable asumir la capacidad que tiene este ciberespacio popular y universal para difundir, de manera veloz, la información que en él se contiene. Del mismo modo, en algunas ocasiones, podrá causar daño a los usuarios. Por lo tanto, debería servir como excusa para que las organizaciones o los Estados definan un mundo virtual distinto (Miró-Llinares, 2011).

2.1.2.3 El ciberespacio abierto y sujeto a revolución permanente

El ciberespacio está sujeto a una revolución constante y a un continuo cambio, debido al uso permanente por parte de la mayoría de la población mundial. De este modo, se creó la Web 1.0 principalmente de lectura. Los sitios web incluían páginas htlm estáticas que se actualizaban con poca frecuencia. El objetivo principal de los sitios web era publicar la información para cualquier persona, en cualquier momento (Aghaei et al., 2012). Sin embargo, era imposible interactuar con el sitio web, ya que su naturaleza era muy pasiva (Choundhury, 2014). Es decir, la primera generación de web permitía buscar información y leerla, pero la interacción con el usuario era más bien escasa.

La Web 1.0 dio paso a la Web 2.0, también conocida como la web de la sabiduría, centrada en las personas, participativa, de lectura y escritura (Aghaei et al., 2012). En esta web los usuarios tienen más interacción, facilitando las prácticas participativas, colaborativas y distribuidas que permiten que las actividades diarias se desarrollen en la web (Choundhury, 2014). Esta web tiene un diseño más flexible, con actualizaciones, creación de contenido y colaboración entre los usuarios.

El término Web 3.0 fue acuñado, por primera vez, por John Markokk del New York Times, al sugerirla como la tercera generación de la web en 2006 (Spivack, 2011). Esta nueva versión, conocida como la web semántica, intenta vincular, integrar y analizar los datos de varios conjuntos de fuentes diversas, para obtener un nuevo flujo de información (Aghaei et al., 2012). Además, es capaz de mejorar la gestión de datos, apoyar la accesibilidad de Internet móvil, simular la creatividad y la innovación, fomentar los fenómenos de globalización, mejorar la satisfacción de los clientes y ayudar a organizar la colaboración en la web social (Choundhury, 2014). El principal objetivo de la web 3.0 es hacer que sea legible, no sólo por los humanos, sino también

por las máquinas. Actualmente, esta web permite conectar toda la información que introducimos en las aplicaciones, en las redes sociales, etc. de una forma más desarrollada que en la versión anterior.

Siguiendo esa revolución permanente del ciberespacio, lo que propone la nueva web del futuro, la Web 4.0, es la interacción entre humanos y máquinas en simbiosis. Es decir, las máquinas serían inteligentes en la lectura de los contenidos de la web y, en función de ello, reaccionarían consecuentemente en la forma de ejecutar y decidir qué ejecutar primero (Choundhury, 2014), implicando una red masiva de interacciones inteligentes. En otras palabras, esta nueva versión ofrecería soluciones a partir de la información que se le aportara y de la existente en la Web.

En definitiva, la web ha ido evolucionando de manera meteórica desde 1989, y se está dirigiendo hacia el uso de técnicas artificiales inteligentes para ser una red masiva de interacciones altamente inteligentes en un futuro cercano (Aghaei et al., 2012; Choundhury, 2014).

#### 2.1.2.4 El anonimato en el ciberespacio

El anonimato es una característica definitoria del ciberespacio, ya que es imposible identificar en él a todos los usuarios del mismo. Existe gran variedad de investigaciones relacionadas con el anonimato en Internet, en las que se han empleado encuestas y entrevistas como sistema de medición (Chesney & Su, 2010; Jessup et al., 1990; Lelkes et al., 2012; Postmes et al., 2013; Thompsen & Ahn, 1992). También se han realizado estudios basados en el anonimato en blogs y foros en determinadas webs (Goga et al., 2013; Peddinti et al., 2014). Los resultados han sido muy variados, creando un debate sobre los beneficios que puede aportar el anonimato al usuario que se siente libre de expresar sus deseos, necesidades o inquietudes, sin miedo a que le juzguen (Correa et al., 2015). O sobre sus inconvenientes, ya que conduce a respuestas caracterizadas por

una mayor hostilidad, un aumento de la agresión y un menor autocontrol (Joinson, 1999). En este sentido, el ciberespacio ofrece a todos los usuarios la oportunidad de expresar sus opiniones libremente (Assimakopoulos et al., 2017), sin importar que sean la difusión de insultos, opiniones extremas, conductas de ciberacoso o discurso de odio (López & López, 2017).

El grado de anonimato que ofrece el ciberespacio, es una de las características que lo hace más atractivo para los criminales (McGrath & Casey, 2002). Es este atributo el que origina que el usuario pierda las inhibiciones y restricciones intrínsecas de la convivencia social (Armstrong & Forde, 2003; Finn, 2004; McGrath & Casey, 2002). Y aumente la sensación de impunidad ante el delito, al separar sus hechos anónimos realizados en el mundo virtual, de su comportamiento en el espacio físico (Suler, 2004). En este sentido, investigaciones como la de Amstrong y Forde (2003) en relación con los pedófilos, o la de Finn (2004) sobre los acosadores, han mostrado que los entornos en línea pueden promover una falsa sensación de intimidad, así como un comportamiento desinhibido, que aumenta la asunción de riesgos y la realización de comportamientos antisociales por parte de un mayor número de personas.

Este anonimato, junto con el resto de características extrínsecas e intrínsecas de Internet, ha fomentado que el ciberespacio se haya convertido, desde hace varias décadas, en un "nuevo lugar" de oportunidad delictiva (Grabosky, 2007).

### *2.2 Discusión académica sobre la oportunidad delictiva en el ciberespacio*

Las teorías de la oportunidad delictiva y en concreto el enfoque de las actividades cotidianas inciden en la existencia de la oportunidad criminal como factor fundamental para que se produzca un hecho delictivo. Siempre y cuando concurra en el tiempo y en

el espacio con un delincuente potencial, un objetivo adecuado y la ausencia de un guardián capaz (Cohen & Felson, 1979). Este enfoque fue el resultado del análisis que realizaron los autores del aumento de la criminalidad en EE. UU., en los años posteriores a la Segunda Guerra Mundial, cuando nadie se podía explicar dicho fenómeno, dado que el país había experimentado una gran prosperidad económica y social. La argumentación, aunque sencilla, fue muy innovadora ya que no se centraron en aspectos psicológicos, genéticos o sociales de los individuos sino en el aumento de oportunidades delictivas gracias al cambio social de las actividades cotidianas de los ciudadanos. Como, por ejemplo, los avances tecnológicos, el uso de cajeros automáticos y la inclusión de la mujer en el ámbito laboral.

Al igual que Cohen y Felson analizaron los cambios sociales en los años setenta del siglo pasado para saber si habían aumentado las oportunidades criminales, en la actualidad, diferentes autores (entre otros, Brady et al., 2016; Grabosky, 2001; Miró-Llinares, 2011; Yar, 2005) dedican sus investigaciones a la transformación relacionada con la manera en la que los individuos interactúan en el ciberespacio. Esta expansión tecnológica y la creciente dependencia de Internet han modificado la forma en la que los individuos interactúan entre sí, realizan negocios, mantienen relaciones con otros y recopilan y difunden información. La tecnología móvil e Internet son un recurso inestimable, pero también constituyen un medio para los delincuentes al ofrecerles nuevas oportunidades de criminalidad (Brady et al., 2016).

Brady y sus colaboradores (2016) sostienen que los cambios en la tecnología desde el uso generalizado de Internet han creado diferentes patrones de "vida cotidiana" que remodelan las estructuras de oportunidades delictivas (Felson & Boba, 2010) en el ciberespacio. El ciberespacio comprende un entorno nuevo, desterritorializado, desmaterializado y desencarnado, que de manera decisiva es discontinuo con el espacio

físico (Capeller, 2001). En este sentido, Capeller (2001) recomienda que la comunidad científica revise los supuestos filosóficos, históricos, sociológicos y por supuesto, los relacionados con el análisis del crimen. A diferencia de Capeller, Peter Grabosky (2001) en su artículo "*Virtual Criminality: Old wine in new bottles?"* sugiere que la cibercriminalidad es básicamente la misma criminalidad que se lleva a cabo en el espacio físico. Para argumentar su afirmación recurre a la teoría de las actividades cotidianas en la que busca establecer una congruencia entre el cibercrimen y el delito en el mundo físico.

Por tanto, para algunos autores (Capeller, 2001; Reyns, 2017; Yar, 2005) la ciberdelincuencia es un fenómeno novedoso en virtud del nuevo espacio en el que se configura. Sin embargo, para otros autores como Grabosky (2001)simplemente se trata de los mismos delitos aunque en entornos distintos. Entre estas dos posturas existe una posición intermedia, como la desarrollada por Fernando Miró en su publicación del 2012 "El cibercrimen: Fenomenología y criminología de la delincuencia en el ciberespacio" en la que afirma que "la cibercriminalidad comparte con la delincuencia todos los elementos definitorios del concepto "crimen", pero dándose los mismos de una forma tal en el nuevo ámbito que es el ciberespacio, que puede influir significativamente en la explicación del delito, y por tanto, en su prevención" (Miró-Llinares, 2012; p.144). Utilizando el símil de Grabosky "*Old wine in new bottles*", Miró-Llinares dice "es vino, pero se bebe de otra forma" (Miró-Llinares, 2012; p.144).

En este debate, el enfoque de las actividades cotidianas, así como la prevención situacional de la delincuencia, intentan mostrar que la ciberdelincuencia puede entenderse y explicarse recurriendo a los recursos aportados por la criminología (Leukfeldt & Yar, 2016). Es una teoría que se ha empleado para analizar diferentes modalidades de crimen, incluyendo robos (Cohen & Felson, 1979), homicidios

(Messner & Tardiff, 1985), robo de vehículos (Rice & Smith, 2002) así como la violencia doméstica (Mannon, 1997).

En este sentido, Yar (2005) ha estudiado la capacidad explicativa de los patrones de la ciberdelincuencia, analizando los elementos principales de la teoría (delincuentes potenciales, objetivos adecuados y ausencia de guardián capaz) y adaptándolos al entorno virtual. Además, como argumentan Yar y Leukfeldt (2016), tiene una estructura analítica muy clara, proporcionando una aplicación, aparentemente sencilla, en casi toda clase de escenarios criminales posibles, permitiendo crear políticas preventivas de la delincuencia como observamos en las estrategias de la "prevención situacional de la delincuencia".

Del mismo modo, Miró-Llinares (2011) reflexiona sobre el momento en el que se formuló el enfoque de las actividades cotidianas, en el que los cambios tecnológicos y sociales modificaron el fenómeno criminal y afirma que, actualmente coexisten en el tiempo la realidad física con un nuevo contexto de oportunidad criminal. Por lo tanto, esta teoría que incide en la relación de los avances tecnológicos y el cambio del delito es especialmente adecuada para explorar cómo operan en el ciberespacio los componentes centrales del triángulo del crimen (García-Guilabert, 2014; Holt & Bossler, 2008; Miró-Llinares, 2011; Pratt et al., 2010; Reyns, 2013; Reyns et al., 2011; Yar, 2005).

Respecto al primer elemento analizado, el agresor potencial, Grabosky (2007) afirma de modo contundente que el número de individuos potenciales para cometer un crimen es un reflejo de la cantidad de personas con acceso a Internet, por lo tanto, cuantos más usuarios estén conectados, más individuos estarán capacitados para emplear esa tecnología para fines ilícitos. Por su parte Yar (2016; 2005) comenta que no parece que haya escasez de delincuentes en el entorno virtual y de manera diversa (estafadores, hackers, piratas informáticos, acosadores, etc.), siendo su presencia fundamental para la

comisión del crimen. En este sentido Miró-Llinares es más explícito (2011) advirtiendo que la oportunidad del delincuente potencial es mucho mayor en el ciberespacio debido a la inexistencia de barreras físicas que impiden el acercamiento entre los delincuentes potenciales y las posibles víctimas.

Asimismo, la arquitectura del nuevo entorno afecta a la motivación del delincuente potencial ya que se elimina la distancia y el tiempo necesarios para realizar un buen ataque, aumentando el número de las posibles víctimas y reduciendo los costes espaciotemporales para cometer los delitos, tanto en la aproximación hacia el objetivo adecuado, como en la huida. Y en su identificación, gracias al anonimato que ofrece el ciberespacio. Esta última idea la comparten la mayoría de los autores, Grabosky (2001) habla de las facilidades que presenta Internet para ocultar la identidad real de los usuarios, que se puede emplear por parte de la policía para encontrar a criminales, pero también permite a los delincuentes ocultar su verdadera identidad haciendo más difícil su captura. El anonimato proporciona a los delincuentes potenciales cierta sensación de impunidad en el ciberespacio (Armstrong & Forde, 2003; Finn, 2004; Suler, 2004), a la vez que le permite reinventarse y adoptar nuevos personajes virtuales alejados de su identidad en el espacio físico (Yar, 2005). Además, el anonimato también puede afectar directamente en la toma de decisiones de los criminales (Miró- Llinares & Johnson, 2018).

Respecto al segundo elemento, los objetivos adecuados, Grabosky (2007) entiende que la distribución de las posibles víctimas estará en función del ascenso de las tecnologías, sin profundizar mucho más en ello. Por su parte, Yar (2016; 2005), Miró-Llinares (2011) y Yucedal (2010), toman como base la conceptualización de los objetivos adecuados en la teoría de las actividades cotidianas, recogidos en el acrónimo VIVA (valor, inercia, visibilidad y accesibilidad).

- Valor: Para Yar (2016; 2005), los objetivos del cibercrimen, al igual que sucede en el espacio físico, varían mucho en las diferentes valoraciones, siendo probable que éstas afecten a la idoneidad del objetivo cuando se considere desde el punto de vista de un posible delincuente, como puede suceder en la piratería informática o en el delito de fraude. Miró-Llinares (2011) comparte la opinión de Yar y argumenta que "en el ciberespacio se da la particularidad de que cosas con poco valor por sí mismas pueden adquirir un valor muy importante gracias a la facilidad para obtener información, relacionarla con la obtenida y convertirla en un objeto de riesgo " (Miró-Llinares, 2012: p.183). Por lo tanto, es un elemento que se debe tener presente para determinar si el objetivo, en el ciberespacio, es adecuado.

- Inercia: Para Yar (2016; 2005) la adaptación de esta característica al ciberespacio es muy complicada, ya que los objetivos en este contexto no poseen propiedades físicas de volumen y masa. Aunque según el autor este elemento se podría adaptar al entorno virtual determinando; por un lado, el volumen de los datos (tamaño de los archivos) que afecta la portabilidad del objetivo y por el otro, el tipo de herramienta (sistema informático) empleada por el delincuente que pone límites a la apropiación de grandes objetivos informativos. Sin embargo, Miró-Llinares (2012) y Yucedal (2010) no siguen esta línea, al plantear que, a excepción de casos muy concretos, los objetivos en el ciberespacio no se distinguen entre sí por sus condiciones extrínsecas. Como indica Yucedal (2010), modificar, eliminar o reproducir la información no tiene gastos en el ciberespacio. De todos modos, los autores comparten la idea de no incluir la inercia como una de las características fundamentales que valoren la idoneidad del objetivo.

- Visibilidad: Los bienes y las personas que son más visibles tienen más probabilidades de convertirse en objetivos adecuados. Yar (2005) argumenta que la conceptualización de la visibilidad en el ciberespacio plantea un problema difícil, ya que Internet es un medio eminentemente público, donde todos los usuarios son visibles. Asimismo, la víctima se expone con determinadas actividades virtuales ante los delincuentes potenciales. Sin embargo, Miró-Llinares (2012) no comparte esta opinión, ya que la inexistencia de barreras en las distancias y el acercamiento de todos los usuarios en un mismo lugar, así como la característica pública que caracteriza al ciberespacio, no consigue que todos los objetivos sean visibles. Serán visibles cuando los usuarios interactúen con otros individuos y con otros servicios. Es decir, no basta con acceder al ciberespacio para ser visible, ya que la cantidad de usuarios que existe en el mundo hace casi imposible su identificación, a menos que exista una interacción entre ellos y en sus actividades (Miró-Llinares, 2012).
- Accesibilidad: Esta característica hace referencia a la capacidad que tiene un delincuente para llegar a su objetivo y luego alejarse de la escena del crimen (Felson, 1998). En el ciberespacio es posible saltar de un punto a otro únicamente con un "click". Por lo tanto, para Yar (2005) parece complicado concebir objetivos distintos en función de la probabilidad de que un posible delincuente tenga acceso a ellos de esta manera. Aunque, en un intento de adaptar el concepto, Yar (2005) lo compara con las contraseñas y otras medidas de autentificación que restringen el acceso a los sitios donde se almacenan objetivos vulnerables, como ordenadores, cuentas de correos, acceso al banco, etc. Pero como ocurre en el espacio físico (con el uso de ganzúas, palancas etc.), estas medidas pueden eludirse a través de herramientas de desencriptación

(Furnell, 2002). Por su parte, Miró-Llinares (2011) argumenta que todos los objetivos son potencialmente accesibles, teniendo en cuenta la contracción de las distancias en el ciberespacio. Por lo tanto, este atributo está relacionado con la capacidad del agresor para llegar hasta él, no a las características del objetivo. Es decir, lo fundamental será la habilidad del delincuente potencial para contactar con un objetivo y llevárselo de la escena del delito.

En resumen, Yar (2016; 2005) plantea que las características del VIVA varían en cuestión de grado, no de tipo. Existiendo similitudes en el valor, pero diferencias significativas en cuanto a la inercia, la visibilidad y la accesibilidad en los objetivos adecuados del ciberespacio, por lo tanto, plantea una reformulación de los mismos. Al igual que Yar, Miró-Llinares (2011) argumenta la posibilidad de mantener el valor como característica imprescindible, para que el objetivo sea adecuado, aunque diferenciándose dependiendo del cibercrimen. Además, plantea una reconceptualización del término visibilidad, denominando "Interacción" a la capacidad de hacer visible a un objetivo en el ciberespacio a través de su movimiento. Sin embargo, suprimiría las características de inercia y accesibilidad (García-Guilabert, 2014).

De esta forma, Miró-Llinares (2012) sugiere una modificación del acrónimo VIVA al IVI (Introducción, Valor e Interacción), conservando el valor del objetivo e incorporando los términos "introducción" para referirse a la acción de desplazar bienes del espacio físico al ciberespacio e "interacción", como características esenciales para conseguir la adecuación del objetivo en el ciberespacio.

Respecto al tercer y último elemento, el guardián capaz, Grabosky (2007) incide en que la tutela puede ser ejercida por individuos o por medios tecnológicos. Además, da por supuesto que tanto los individuos como las organizaciones que posean objetivos adecuados que quieran proteger, tendrán un administrador de sistemas y habrán

invertido en ciberseguridad, ya que no es posible ubicar a un agente de policía al lado de cada ordenador. Por lo tanto, la prevención de la ciberdelincuencia es una tarea que se debe compartir por los gobiernos, los ciudadanos y las instituciones de la sociedad civil por igual (Grabosky, 2007). Dicho esto, el autor entiende que en última instancia debe ser la propia víctima la que con sus acciones evite, desde el comienzo, que el delincuente consiga su propósito (Grabosky, 2001). En esta misma línea, Leukfeldt y Yar (2016) plantean la opción de que los usuarios pueden hacerse menos accesibles a los delincuentes tomando medidas de protección (instalando antivirus y manteniéndolos actualizados). Estaríamos ante el guardián capaz técnico. Por lo tanto, según los autores (Leukfeldt & Yar, 2016) los usuarios de Internet con un alto grado de conocimiento técnico y/o que son conscientes de los riesgos a los que se enfrentan en el ciberespacio son más capaces de anticiparse a los ataques y, en consecuencia, tienen un menor riesgo de convertirse en víctimas (tutela personal capaz). Como el resto de autores, Miró-Llinares (2011) realiza un análisis del guardián capaz como encargado de la vigilancia de los objetivos adecuados. Y, como ya indicaron otros autores (Choi, 2008; Peter Grabosky, 2007; Holt & Bossler, 2008; Yar, 2005) los guardianes pueden ser los distintos programas creados para evitar las intrusiones de los delincuentes, como los antivirus, los programas antiespías, *firewall*, etc. Pero estos elementos, como indica Miró-Llinares (2011), dependerán de la actuación de la víctima, siendo fundamental la propia autodefensa. Estos autores están de acuerdo en la disminución de la capacidad del guardián capaz en el ciberespacio para evitar la comisión de un delito, entendido como lo formularon Cohen y Felson (1979). Sin embargo, en el ciberespacio gana protagonismo la víctima ya que sus acciones de autodefensa serán fundamentales para la prevención de los posibles cibercrímenes.

En definitiva, la teoría de las actividades cotidianas defiende que, para que se lleve a cabo un delito deben confluir en el espacio y en el tiempo tres elementos principales: delincuente potencial, objetivo adecuado y ausencia de un guardián capaz (Felson, 1998). Pero, la incapacidad de trasladar al ciberespacio los postulados de esta teoría de "convergencia en el espacio y el tiempo" hace que sea complicada la explicación de la génesis de los ciberdelitos (Yar, 2005). Por lo tanto, tendremos que analizar las nuevas características del ciberespacio para, posteriormente, investigar de qué manera estos cambios han afectado a los elementos del crimen.

# Capítulo 4

# LOS CIBERLUGARES Y LA DETECCIÓN DE LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO EN TWITTER

"The Internet is becoming the town square of the global village of tomorrow"

(Bill Gates, 2013)

## 1. El Ciberlugar como Evolución de la Aplicación de la Criminología Ambiental al Ciberespacio

El ciberespacio, entendido como un espacio virtual que, no siendo un lugar geográfico, se asemeja a él al ser posible la realización de acciones como en el espacio físico (Miró-Llinares & Johnson, 2018), ha modificado la forma de interaccionar de las personas gracias a la contracción total de las distancias, y el favorecimiento de la comunicación entre ellas (Miró-Llinares, 2011).

### *1.1. El concepto de ciberlugar. Sobre la idea de convergencia digital*

El análisis del crimen y el lugar desde la perspectiva de la teoría de las actividades cotidianas siempre ha considerado que la convergencia ocurre en algún momento y lugar. Desde este punto de vista, el lugar ha sido tradicionalmente algo dinámico, pero siempre físico. Como por ejemplo un parque, una zona de ocio, una dirección específica, etc. (Brady et al., 2016). Con el desarrollo de Internet, sin embargo, el "lugar" es ahora accesible para todos, por lo tanto, esta convergencia física no es un

requisito de la ciberdelincuencia. Pero, a pesar de ello, el espacio geográfico no es irrelevante (Miró-Llinares & Johnson, 2018).

En el ciberespacio, los usuarios convergen en el mismo lugar, ya sea todo el tiempo o a intervalos (Reyns et al., 2011). De este modo, a medida que la población aumenta sus actividades *online*, también aumentará la posibilidad de convertirse en víctima. Por ello, el concepto de "ciberlugar" debe especificarse todavía más, a medida que se sigue aplicando la teoría de las actividades cotidianas a la ciberdelincuencia. Aunque esto también dependerá de la naturaleza del delito (Brady et al., 2016). Como en los delitos de acoso, que pueden suceder en tiempo real, o después de una interacción de acciones por parte del delincuente (amenazas, requerimientos sexuales, etc.) y la recepción de esas acciones por parte de la víctima.

Tradicionalmente, el concepto de lugar se ha utilizado para referirse a lugares concretos en el espacio físico. No obstante, también se puede utilizar para referirse a nodos o áreas de actividad en Internet en las que no se está físicamente ubicado, pero donde sí se puede actuar (Miró-Llinares & Johnson, 2018). Como se ha indicado, el ciberespacio no es uniforme, pero permite la interacción entre los usuarios. Y entre las personas y la información en diferentes modalidades, como por ejemplo los correos electrónicos, las páginas web o las redes sociales. En este sentido, como señalan Miró-Llinares y Johnson (2018), cada uno de ellos podrían considerarse diferentes ciberlugares. Las cuentas de correo electrónico permiten la interacción entre los usuarios que se encuentran geográficamente separados; las páginas web son lugares que se visitan y las redes sociales son lugares específicos en los que los usuarios, o su representación, puede permanecer una cierta cantidad de tiempo e interactuar con otros.

Estos y otros espacios pueden configurarse de muchas maneras diferentes, y pueden ser estructurados por muchos lugares diferentes. La manera de identificar los distintos

ciberlugares es la de considerar las diferentes maneras de comunicación, o contacto, que tienen lugar en ellos. No es lo mismo la interacción con una página web y la que se produce en una red social (Miró-Llinares & Johnson, 2018).

Por lo tanto, esta diferencia afectará la forma específica que pueden tomar los eventos criminales y cómo se producen. Según Miró-Llinares y Johnson (2018), los factores que influyen en el hecho de que un delincuente se encuentre con un objetivo adecuado en ausencia de un guardián capaz en lugares virtuales son los siguientes:

- La modalidad de contacto facilitada por la tecnología utilizada.
- El posible potencial para la observación externa y la gestión de lugares y, por lo tanto, la mayor o menor capacidad de tutela.
- El tipo de actividad en el ciberespacio que sea habitual para ese lugar virtual concreto en cuestión.

De este modo, si efectivamente el "lugar" en el triángulo del crimen es una dirección web, los proveedores de servicios de Internet y/o los desarrolladores de software pueden ser considerados en última instancia como administradores de lugares (Henson & Reyns, 2015; Sampson et al., 2010). Estos administradores de lugares tendrán una responsabilidad adicional de desarrollar protocolos efectivos y basados en la evidencia para garantizar la seguridad y protección general (Brady et al., 2016).

### *1.2. Lugares para la concurrencia en el ciberespacio*

En el ciberespacio puede haber diferentes tipos de interacción. Por lo tanto, los ciberlugares serán distintos en función de si la comunicación es sincrónica (es decir, en tiempo real, como sucede en plataformas como Google Hangouts, Skype, Facetime, etc.), o existe una demora entre la transmisión y la recepción de la información (como ocurre con el correo electrónico, los foros, salas de chat, etc.) (Miró-Llinares &

Johnson, 2018). También existen espacios virtuales que permiten todo tipo de comunicación, como por ejemplo la red social Twitter, en la que el contacto es asincrónico, pero tiene una aplicación, Periscope, que facilita la comunicación en streaming.

Los lugares en Internet se caracterizan, entre otras cosas, por la manera de interactuar entre sus usuarios. Por lo tanto, los cibercrímenes se desarrollarán en función de los mismos. Algunos cibercrímenes se cometerán en ciberlugares donde el contacto es asincrónico, como la transmisión de virus a través de correo electrónico. En este caso, se envía un mensaje que contiene un archivo adjunto malicioso, pero no se activa hasta que el usuario abra el archivo adjunto. Por el contrario, otros cibercrímenes se producirán cuando el contacto sea en tiempo real, como es el caso de la petición de actos sexuales a un menor (Miró-Llinares & Johnson, 2018). También existe la posibilidad de combinar ambas modalidades. Como, por ejemplo, una serie de envíos de correos electrónicos entre la víctima y el delincuente, que llevan finalmente a un acoso a través de Hangouts.

Por otro lado, la tutela y la vigilancia natural de los ciberlugares es complicada, aunque nos encontremos ante un espacio público como es Internet. Del mismo modo que ocurre en el espacio físico, como argumenta Katyal (2003) las características de los ciberlugares facilitan y limitan las oportunidades de la vigilancia natural sobre las oportunidades delictivas.

Por su parte, Miró y Johnson (2018) enumeran tres factores asociados con la configuración de los ciberlugares que influyen en la vigilancia natural:

- Restricciones de acceso a los ciberespacios, que pueden variar desde el acceso totalmente restringido (privados) hasta la entrada abierta a todos los lugares de los usuarios (públicos).
- El volumen de tráfico en cada lugar.
- La superposición de la red subyacente de cada red, ya que las denominadas redes superpuestas dificultan la localización del indicador IP de los usuarios que viajan en ellos, ocultando así la identidad de los usuarios.

Por ejemplo, las organizaciones terroristas pueden aprovechar la escasez de vigilancia natural en determinados ciberlugares para reclutar objetivos adecuados (Mahmood & Rohail, 2012), como salas de chat privadas de difícil acceso que tengan poco tráfico.

Sin embargo, otros lugares como Facebook pueden permitir una vigilancia activa por parte de los “amigos” del usuario. De este modo, estas plataformas ofrecen más oportunidades para la observación externa y la "vigilancia natural" y, por lo tanto, un mayor potencial de tutela (Miró-Llinares & Johnson, 2018).

Como en el espacio físico, las personas se dedican a determinadas actividades en determinados ciberlugares en determinados momentos. Por lo tanto, lo más probable es que los diferentes objetivos adecuados buscados por los criminales se encuentren en ciberlugares específicos y en momentos concretos (Miró-Llinares & Johnson, 2018). En este sentido, en Internet la diversidad de actividades es amplia. Además de ser un lugar de trabajo, también podemos encontrar actividades para el ocio, el consumo, la información, la banca *online*, etc. Y al igual que la actividad en el mundo físico, la actividad que se realiza en el ciberespacio tiene un ritmo, y unos patrones de conducta (Li et al., 2013).

Y, como ocurre en el espacio físico, los ciberlugares son administrados por gestores. Estos responsables pueden prestar más o menos atención a estos lugares, pero del mismo modo que el dueño de un bar debe velar por la seguridad de su local, los gestores del ciberespacio deben hacer lo mismo (Miró-Llinares & Johnson, 2018). Por ejemplo, en muchos sitios web se incluye la publicidad *online*. Mientras que la mayoría de la publicidad puede ser legítima, otra parte no lo es. Se sabe que estas prácticas de gestión varían considerablemente (e.g, Zarras et al., 2014) y pueden hacer que los usuarios sean más o menos vulnerables a los ataques.

Así, como argumentan Miró-Llinares y Johnson (2018), todos estos caracteres definen las características ambientales de los diferentes ciberespacios, de manera que esta composición influirá en la oportunidad delictiva relacionada con ellos.

### *1.3. Patrones delictivos y ciberlugares*

Una vez definidos los lugares virtuales del ciberespacio como ciberlugares en los que es posible la convergencia entre los objetivos adecuados y los delincuentes potenciales. Y una vez diferenciados y categorizados, debemos analizar la relación entre ellos y los delitos a partir de las premisas, ya conocidas, de la criminología ambiental (Miró-Llinares & Johnson, 2018). En este sentido, ya sabemos que el crimen no se distribuye de manera aleatoria en el espacio y en el tiempo, sino que mantiene unos patrones y se concentra en los denominados *hots spots* (Brantingham & Brantingham, 1991).

Al igual que ocurre en el espacio físico, en el ciberespacio la delincuencia también se concentra en determinados lugares virtuales, en función de sus características. Por ejemplo, como cita Miró-Llinares (2015), será más común que a través del correo electrónico se produzcan infecciones de malware o robos de información confidencial (phishing). Por otro lado, Kwan y Skoric (2013) argumentan que el ciberacoso parece

ser más común en las redes sociales: en Facebook, y en los juegos *online*, la victimización realizada con insultos y amenazas, y a través de mensajes privados directos, es más frecuente que en los muros públicos o semipúblicos; y en Twitter, el odio y la radicalización (Miró-Llinares & Johnson, 2018). Por su parte, Hartel y sus colegas (2011) exponen que el acoso infantil es más común en las salas de chat que en otros ciberlugares.

En definitiva, determinados ciberlugares son más susceptibles de amparar algunos tipos de cibercrímenes debido al tipo de actividad que se desarrolla en los mismos. Del mismo modo, los delitos cometidos en cada uno de estos ciberlugares tampoco se distribuirán de la misma forma (Miró-Llinares & Johnson, 2018).

Si hablamos de ciberlugares donde se concentra la actividad delictiva, debemos mencionar de manera especial pero muy sucinta, la *Dark web* como parte de la *Deep web*, la cual no está indexada por los motores de búsqueda estándar. La Dark web está alojada en servidores que utilizan algoritmos de encriptación. Además, para acceder a ella se debe utilizar un navegador especial como, por ejemplo, la red Tor que utiliza la misma encriptación (Miró-Llinares & Johnson, 2018). El acceso a la Dark web está restringido, lo que hace muy complicada su tutela, fomentando las ofertas de pornografía infantil, drogas y otros productos ilegales (Moore & Rid, 2016).

Existe otro principio fundamental de criminología ambiental válido para el ciberespacio:  el que relaciona las concentraciones de eventos criminales con las actividades cotidianas de los usuarios, y que los lleva a convertirse en generadores del crimen (Brantingham & Brantingham, 1995). Es decir, es el uso de la banca electrónica lo que lleva al delincuente a buscar el acceso a las claves, o las actividades de los menores en las redes sociales las que convierten a estos lugares en propicios para la violación de su privacidad, etc.

Investigaciones como las de Holt y Bossler (2009), Hutchings y Hayes (2009) y Ngo y Paternoster (2011), han identificado los factores de riesgo de victimización por diferentes ciberdelitos relacionados con actividades cotidianas y el uso de las TIC.

Teniendo en cuenta todo lo expuesto, la prevención situacional del delito, con su énfasis en la modificación de las estructuras de oportunidad, puede ser la mejor opción para su prevención (Clarke, 1983). Los desarrollos teóricos que se han mencionado sobre los ciberlugares son importantes, ya que las dinámicas de la prevención situacional de la delincuencia pueden ser aplicadas para prevenir determinados delitos en el ciberespacio, concretamente en las redes sociales.

## 2. Aplicación de la Teoría de los Ciberlugares para la Detección de la Comunicación Violenta y el Discurso en las redes sociales

Los ciberlugares suelen ser utilizados de manera incorrecta para difundir contenidos abusivos, degradantes o perjudiciales para la sociedad (Mossie & Wang, 2018). La comunicación violenta y el discurso de odio se ha convertido en un problema significativo para todo tipo de plataformas *online* en la que el contenido creado por los usuarios aparecen desde secciones de comentarios de los sitios web de noticias hasta las sesiones de chat en tiempo real (Mossie & Wang, 2018). En este sentido, las redes sociales permiten que cualquier usuario con una conexión a Internet pueda difundir cualquier tipo de mensajes llegando a millones de personas (Silva et al., 2016) de manera casi instantánea.

Teniendo en cuenta estas circunstancias, el objetivo de este apartado es analizar la comunicación violenta y el discurso de odio que se difunde en las redes sociales en

general y en Twitter, en particular, a través de los enfoques de la criminología ambiental centrada en el concepto del ciberlugar.

Como ya se ha indicado anteriormente, del mismo modo que los delitos tradicionales, los ciberdelitos se producen con mayor frecuencia en algunos lugares más que en otros y en determinados momentos. Así, las redes sociales contienen todos los elementos adecuados para convertirse en el ciberlugar preferido de los usuarios para difundir mensajes violentos y de odio.

En este sentido es importante recuperar los elementos que desarrollaron Miró-Llinares y Johnson (2018) para identificar las características que tienen las redes sociales como ciberlugares adecuados para la difusión de contenido violento:

1. La modalidad de contacto facilitada por la tecnología utilizada. Es decir, el tipo de contacto existente en las redes sociales ya sea secuencial o a tiempo real.

   En el ciberespacio en general y en las redes sociales en particular, existen diferentes formas de interactuar. Las redes sociales más destacadas permiten cualquier tipo de comunicación, aunque se caractericen por su contacto asincrónico. Por ejemplo, YouTube se identifica como la plataforma más icónica donde colgar vídeos de los usuarios, pero también permite realizar conexiones en directo, al igual que Instagram y Facebook. Por su parte, Twitter, comparándolo con otras redes sociales es un entorno de comunicación masiva, interactiva e inmediata de contenidos. Aunque permite la comunicación en *streaming* (a través de su aplicación Periscopio) y los mensajes directos a usuarios concretos fuera de la vista del resto de la red, Twitter funciona esencialmente como una plaza pública en la que se utiliza la comunicación almacenada y reenviada para expresar contenidos que pueden ser observados y compartidos por un gran número de personas

(Marwick & Boyd, 2011). Si añadimos que la comunicación política o ideológica se ha hecho cada vez más frecuente en Twitter (Bode & Dalrymple, 2016), parece comprensible que esta red social se utilice comúnmente para difundir la comunicación violenta y el discurso del odio (Schmidt & Wiegand, 2017).

2. El posible potencial para la observación externa y la gestión de lugares y, por lo tanto, la mayor o menor capacidad de tutela.

Este apartado, se basa fundamentalmente en la visibilidad del usuario. La visibilidad, en este caso, produce un efecto distinto al que tradicionalmente se la ha atribuido a la criminología. De hecho, esta característica siempre se ha considerado un elemento negativo desde la perspectiva del agresor, ya que para consumar de manera adecuada un hecho delictivo en el espacio físico, una de las principales características es proteger la identidad del delincuente para no ser identificado. Sin embargo, si un usuario tiene como finalidad difundir su mensaje violento y de odio, elegirá las redes sociales donde tenga más visibilidad, a través de sus seguidores. Del mismo modo cobra mucha importancia en las víctimas ya que, si una víctima no es visible en la red social, no existirá para el resto de los usuarios y será complicado que sea objeto de estos mensajes tan dañinos.

En este sentido los trolls, fingen esforzarse para formar parte de esta comunidad, pero las intenciones reales son las de causar trastornos y aumentar el conflicto, con el único fin de divertirse. De este modo, los usuarios con pensamientos radicales utilizan las redes sociales para fomentar el odio contra determinados individuos o colectivos, causando un efecto contagio, que puede dañar gravemente a determinados usuarios (Del Vigna et al., 2017) En algunos casos, los usuarios con más experiencia se pueden enfrentar a determinadas amenazas y a estos trolls, pero la gran mayoría de ellos no pueden soportar fácilmente los ataques (Del Vigna et

al., 2017), principalmente los menores y los que están expuestos al juicio público por sus condiciones personales (periodistas, políticos, actores, influencers, youtubers, etc.).

En relación con este aspecto, una de las características del cibercrimen que genera patrones delictivos es la que se produce por la victimización repetida. El mecanismo es similar al de la delincuencia tradicional, ya sea debido a delincuentes particularmente reincidentes o a lugares con características criminógenas. Así, se pueden producir altas concentraciones de victimización en las redes sociales, relacionadas con las explicaciones de "impulso o de bandera" (Johnson,2008 citado en Moneva, 2020), en la que determinados entornos tienen características estáticas que los identifican continuamente como vulnerables al delito. En el caso concreto de Twitter sucede que los usuarios tienen sensación de impunidad ya que pueden difundir repetidamente comentarios discriminatorios contra un grupo vulnerable en ausencia de una sanción de los moderadores de la plataforma (Moneva, 2020). Como resultado, el ofensor observa su oportunidad para continuar difundiendo sus mensajes violentos y de odio. Imaginemos que un usuario ofende a una víctima con un comentario y además recibe algún apoyo de otro usuario en forma de otro mensaje de similares características o un retweet: habría conseguido su objetivo. Así, el delincuente percibe una oportunidad de continuar ofendiendo a ese mismo usuario con más mensajes en el futuro. La explicación de bandera hace referencia al entrono en el que se lleva a cabo el delito, el cual posee características estáticas que lo identifica continuamente como vulnerable al delito, en vez de centrarse en el delincuente (Moneva, 2020).

De este modo, en los últimos años, los usuarios no sólo pueden advertir temor ante una posible amenaza en el espacio físico, sino que esa sensación de inseguridad y de miedo al delito se traslada a este ciberlugar, las redes sociales (Vozmediano-Sanz et al., 2009).

3. Tipo de actividad que se desarrolla en las redes sociales.

En el primer capítulo hemos visto las distintas clases de redes sociales que existen. Por un lado se encuentran las redes sociales verticales, caracterizadas por tener una temática en común que facilita la comunicación entre los usuarios que tienen intereses similares (De Haro, 2010) y por el otro, las redes sociales horizontales, que carecen de una temática concreta (De Alsola, 2008). Dentro de las redes sociales verticales, encontramos de todo tipo: profesionales (e.g. Linked In), de citas (e.g. Tinder) o de viaje (e.g. Bla, bla, car), entre otras. Poco se conoce de casos relacionados con la comunicación violenta y el discurso de odio en este tipo de redes, quizás debido a sus características intrínsecas. En ellas, todos los usuarios tienen el mismo objetivo, ya sea laboral, sentimental o de ocio. Por ello, no se configuran como ciberlugares adecuados para fomentar este tipo de comportamientos disruptivos.

Sin embargo, en las redes sociales verticales como YouTube, Instagram, Facebook o Twitter, ocurre todo lo contrario. Son plataformas donde no existe una unión generalizada sobre temas concretos, sino que cada usuario tiene sus propias inquietudes, pensamientos y hobbies que pueden coincidir con unos usuarios, pero con otros no. De ahí que este tipo de plataformas, concretamente Twitter, sí se configuren como ciberlugares apropiados para difundir mensajes relacionados con la comunicación violenta y el discurso de odio.

# 3. La Detección de la Comunicación Violenta y el Discurso de Odio en los Microlugares de Twitter

## *3.1 La Red Social Twitter*

La red social Twitter es un servicio *online* de microblogging que permite a los usuarios estar en contacto a través de mensajes instantáneos de no más de 280 caracteres (en un principio fueron 140) denominados tweets.

Los creadores de Twitter, Jack Dorsey, Biz Stone y Evan Williams tenían como objetivo crear una red social que se basara en mensajes cortos, del estilo de los SMS. Así, el 21 de marzo de 2006, Dorsey hizo público su primer mensaje en una versión de prueba, en una plataforma que denominaron Twttr (Orihuela, 2011). Se lanzó al público en junio de 2006, y rebautizaron la versión final como Twitter. Pero, realmente comenzó a tener gran popularidad en el primer trimestre del 2007, llegando a España en el 2009. Desde ese momento, su crecimiento fue notorio y su transformación también. De hecho, gracias al Interfaz de Programación de Aplicaciones (API) de la plataforma, se ha configurado un grupo de desarrolladores que han lanzado multitud de aplicaciones que optimizan sus funcionalidades (Orihuela, 2011).

En Twitter se puede hablar de cualquier tema de interés, ya que no existe limitación al respecto. Originalmente, la plataforma planteaba una pregunta para fomentar las publicaciones de los usuarios: *What are you doing*? (¿Qué estás haciendo?). A veces la gente respondía bastante bien, contando con todo tipo de detalles lo que habían desayunado, el viaje que estaban realizando o las gestiones del trabajo. Pero es cierto, que a medida que el servicio ha ido ganando usuarios, la gente lo utiliza cada vez más para hablar sobre lo que leen, ven, escuchan, piensan, etc. De hecho, Twitter se ha convertido en un actor clave que distribuye ideas y comentarios sobre lo que le interesa a la gente (O'Reilly & Milstein, 2011). Así, en noviembre de 2009, se sustituyó de

manera acertada la pregunta inicial *What are you doing*" por *What's happening?* (¿Qué está pasando?).

Twitter se ha convertido en una plataforma de comunicación muy poderosa y atractiva que resulta ser muy útil para un sinfín de necesidades personales y profesionales. Publica los mensajes de manera instantánea y sin restricciones (a excepción de las cuentas protegidas) para que cada usuario decida qué perfiles les interesan más. Es tal la extensión del uso de Twitter que, según indican estudios realizados por We are social y Hootsuit, existen 251.000.000 millones de personas en el mundo que utilizan esta plataforma, siendo el 34% mujeres y el 66% hombres. En España, son miembros de esta red social alrededor de 6.000.000 millones de personas, con una distribución por sexo similar.

### 3.1.1 *Características especiales de Twitter*

Algunas de las características que hacen que Twitter sea una red social única son las siguientes (O'Reilly & Milstein, 2011):

#### 3.1.1.1 Limitación de caracteres

Desde su creación, uno de los rasgos más característicos de Twitter, fue la limitación de la extensión de los mensajes, los cuales no podían superar los 140 caracteres. De este modo, la sencillez, brevedad e ingenio de las publicaciones era patente, facilitando su lectura al resto de la comunidad. Además, desafiaba a los usuarios a concentrar sus ideas, reflexiones, comentarios, etc. en pocos caracteres, y sobre cualquier tema que fuera de interés para ellos, ya que la ausencia de predeterminación relacionada con el contenido de los mensajes publicados provoca que la comunicación pueda ser más libre.

En el año 2017 Twitter decidió doblar el espacio, convirtiendo los 140 en 280 caracteres. Desde la red social se argumentó que, en ciertos idiomas como el inglés, francés, español o portugués, era muy complicado expresarse de una manera correcta y

que rápidamente se llegaba al límite, obligando a editar y eliminar palabras que eran importantes para los usuarios, fomentándose así una sensación de frustración. No ocurría lo mismo con otros idiomas como el chino, japonés o coreano, ya que pueden transmitir el doble de información en un carácter (Rosen & Ihara, 2017). Twitter realizó una investigación que confirmó la realidad de esta hipótesis: El porcentaje de tweets enviados en japonés -que tenían 140 caracteres- era sólo de un 0,4%. Sin embargo, en el caso del idioma inglés, el de tweets con 140 caracteres era mucho mayor: un 9% (Rosen & Ihara, 2017). Por lo tanto, estos resultados reforzaron la iniciativa del aumento del número de caracteres, para lograr así el objetivo de que todas las personas, independientemente de su idioma, se pudieran expresar más fácilmente.

#### 3.1.1.2 Conexión temática mediante hashtag (etiqueta)

El término hashtag (etiqueta) se refiere a una palabra clave precedida del signo # (almohadilla) que identifica un tema, evento o asunto al que hace referencia un tweet, ayudando a categorizar los mensajes en la propia red social (O'Reilly & Milstein, 2011).

En Twitter no existe un tema predeterminado del que se deba hablar, sino que dependerá del interés de los usuarios, por ello el recurso del hashtag es una idea sencilla pero eficaz, ya que no existe manera de categorizar la multitud de mensajes que se publican al día. De este modo, cuando alguien quiere agrupar los mensajes que están relacionados con ese tema, crea un término corto y lo precede con el símbolo #. El resto de usuarios únicamente tendrá que redactar el comentario y añadir el hashtag antes de publicar el tweet.

#### 3.1.1.3 Los Retweets (RTs)

Según O'Reilly y Milstein (2011), el término "Retweeting" es uno de los más básicos que existe en Twitter, pero también uno de los más importantes. Hace referencia

a volver a publicar desde tu propia cuenta el tweet de otra persona para darle relevancia. Los RTs favorecen que los mensajes más importantes para los usuarios funcionen en Twitter. Pero, además, es una manera de comunicarse doblemente: por una parte, al creador del tweet original se le indica a través de una mención que su publicación es destacada, y por otra, a los seguidores de la cuenta propia se les comunica que existe un contenido ajeno al que deberían prestarle atención (Orihuela, 2011). En definitiva, como indica Kwak y sus colaboradores (2010), el mecanismo de retweet permite a los usuarios difundir la información seleccionada más allá del alcance de los seguidores del tweet original.

#### 3.1.1.4 Inclusión del contenido audiovisual (enlaces externos)

En Twitter no existe la contextualización de los mensajes. Los dos recursos con los que cuentan los usuarios para delimitar la temática de sus publicaciones son los hashtags y los enlaces de hipertexto. Estos enlaces permiten que el usuario comparta con sus seguidores una referencia que le de mayor sentido a su mensaje. En el uso de ellos es importante reducir el tamaño de la URL para ahorrar espacio y disponer de caracteres en los que se pueda indicar la naturaleza del contenido de la misma, como vídeos, fotos o enlaces externos.

#### 3.1.1.5 Trending Topics (temas de actualidad)

Twitter rastrea las palabras, frases y hashtags que son más frecuentes y los publica bajo el título “Trendig Topics”, en una lista de los diez temas más populares del momento. En una barra lateral derecha de la página principal de cada usuario, de forma predeterminada, a excepción de se modifique la configuración (Kwak et al., 2010). Debido a que Twitter actualiza esta lista regularmente, los temas de moda reflejan las inquietudes de la gente, revelando noticias de última hora, incluso antes de que los principales medios de comunicación empiecen a informar.

#### 3.1.1.6 Las relaciones entre los usuarios son asimétricas

A diferencia de la mayoría del resto de redes sociales (como Facebook), las relaciones entre usuarios son optativas. Es decir, no tiene porqué existir un interés mutuo entre ambos. En la terminología de Twitter, cuando a una persona le interesan los contenidos de otra, se denomina seguimiento. Por lo tanto, el hecho de seguir (follow) y ser seguido no requiere reciprocidad (Kwak et al., 2011). Los seguidores (followers) son personas que reciben las actualizaciones de Twitter de otras personas, sin necesidad de que haya seguimiento mutuo (Huberman et al., 2008). Puede darse el caso en el que el interés desaparezca, y el usuario interrumpa el seguimiento (unfollow), ya sea porque las publicaciones ya no son atractivas para él, o porque está en disconformidad con las mismas.

#### 3.1.1.7 Mensajes directos y menciones

Los usuarios de Twitter pueden realizar publicaciones generales, o a personas concretas, a través de la comunicación privada entre los usuarios. Aunque las actualizaciones directas se utilizan para comunicarse directamente con una persona concreta, son públicas y cualquiera puede verlas (Huberman et al., 2008). Además, se pueden realizar menciones haciendo referencia a otro tuitero en un mensaje utilizando el símbolo @ seguido del nombre del usuario. Por otro lado, también existe la posibilidad de crear cuentas protegidas para que las publicaciones no sean públicas y que sea el usuario el que autorice al resto para acceder a la publicación de sus tweets.

#### 3.1.1.8 Requisitos de registro en Twitter (los metadatos)

Para realizar el registro en Twitter se pueden cumplimentar una serie de requisitos que, como veremos en el siguiente apartado, serán fundamentales para el análisis de la comunicación violenta y el discurso de odio en esta red social:

- Nombre de usuario: El usuario (tuitero) debe identificare con un nombre para ser reconocido dentro de la plataforma (e incluso, en ocasiones, fuera de ella). Esta identidad es aconsejable que sea breve (no superior a 15 caracteres) y que esté relacionada con las características del usuario.

- Biografía (descripción): Es la presentación que muestra el usuario al resto de usuarios de Twitter y en la cual debe describir, en no más de 160 caracteres, la relación de su identidad en el espacio físico con la clase de publicaciones que realizará en su cuenta.

- Foto de perfil: Será la imagen (avatar) del usuario en la red social, en la que debe quedar presente la identidad del tuitero.

- URL externa: Hace referencia al sitio web al que pueden acudir el resto de los usuarios, para obtener mayor información del propietario de la cuenta (enlace a LinkeIn, a la entidad donde trabaja, etc.)

- Vecindario: Indica los usuarios a los que sigue y quiénes son sus seguidores.

- Ubicación del usuario: En este campo, el usuario debe incluir la información de su ubicación. Es decir, la ciudad, provincia o país donde vive.

- Geoposicionamiento del tweet: El usuario puede indicar la ubicación precisa de sus publicaciones, siempre y cuando habilite esta opción, ya que se encuentra desactivada de forma predeterminada.

### *3.2 La influencia de Twitter en la difusión de la comunicación violenta y el discurso de odio*

Al comienzo de este capítulo hemos analizado el concepto de ciberlugar, así como su relación con la comunicación violenta y el discurso de odio, concretamente en las redes sociales. Es indiscutible que las características de las mismas hacen que la difusión de los mensajes ofensivos sea mayor que en el espacio físico, por ello nos vamos a centrar en este fenómeno y concretamente, en su relación con la red social Twitter, objeto de nuestra investigación.

Parece que Twitter, después de la ampliación de sus caracteres, es un lugar algo más educado y en el que abundan los debates y las opiniones algo más constructivas (Jaidka et al., 2019). Los investigadores analizaron más de 350.000 tweets en EEUU entre enero de 2017 y marzo de 2018, identificando el estilo, el contenido, la carga emocional de los mismos y la profundidad del debate. Los resultaron mostraron que tanto los insultos, como los tweets ofensivos, de ira, y los comentarios de mala educación se habían reducido después del aumento de sus mensajes a 280 caracteres. Por otra parte, se habían multiplicado el análisis, la argumentación constructiva y las expresiones educadas. Es cierto que son datos con una perspectiva optimistas, pero lamentablemente Twitter sigue siendo un ciber lugar donde abundan los haters y donde la comunicación violenta y el discurso de odio están patentes.

La propagación de la información, y concretamente del discurso ofensivo, tiene características especiales en Twitter. Debido a que es una plataforma sincrónica, global y con características sociales, facilita la rápida circulación y viralidad de las publicaciones, noticias u opiniones y la multiplicación de las mismas, sobre todo a través de los retweets (Orihuela, 2011). En este sentido, Twitter no es un lugar homogéneo donde todo ocurre de la misma manera en todas partes dentro de él. Es bien

sabido, por ejemplo, que la distribución temporal de los mensajes no se produce de forma aleatoria (Miró-Llinares & Rodríguez-Sala, 2016). Por lo tanto, la mayor parte de este tipo de mensajes se concentrarán en los diferentes microlugares que posean las características más adecuadas para ello. Como, por ejemplo, sucede en el espacio físico donde la posibilidad de que se comentan determinados delitos aumentará en las calles sin iluminación ni vigilancia.

### *3.3 Los microlugares de Twitter*

La comunicación violenta y el discurso de odio, al igual que ocurre con otras conductas en el espacio físico y en el ciberespacio (Miró-Llinares & Johnson, 2018), se distribuye en función de determinados patrones dependiendo de las características de los microlugares en los que se producen (Miró Llinares & Johnson, 2018).

Por lo tanto, si estudiamos la red social Twitter como un ciberlugar deberemos dirigir nuestro análisis concretamente a los microlugares. El estudio de estos microlugares, como lugares donde suelen ocurrir los delitos, sirve para comprender qué tipo de convergencia se produce en el mismo para que el resultado sea un hecho criminal. Es decir, en Twitter se crean microentornos digitales únicos como resultado de la combinación entre las cuentas de los usuarios y lo que dicen en sus publicaciones (tweets) a otras personas (Miró-Llinares et al., 2018).

Es cierto que Twitter permite la comunicación en *streaming* a través de Periscopio, y la privacidad de los mensajes directos a otros usuarios concretos fuera de la comunidad, pero la mayor característica de Twitter es la publicación generalizada de contenidos de manera masiva, inmediatos e interactivos (Miró-Llinares et al., 2018). Es por ello comprensible que se utilice habitualmente para difundir discurso de odio (Schmidt & Wiegand, 2017) teniendo en cuenta que es frecuente que los mensajes tengan un contenido ideológico o político (Bode & Dalrymple, 2016). Y como argumentan

diversos autores (Berger & Morgan, 2015; Vielleux-LePage, 2016; Weimann, 2018), se haya convertido en la red social preferida de los grupos terroristas y extremistas para fomentar la radicalización.

Otra particularidad muy característica de Twitter es la restricción de la longitud de sus publicaciones (en un primer lugar 140 caracteres, actualmente 280), la cual limita las posibilidades de interacción entre los usuarios, centrando la actividad de los incitadores al odio o terroristas en normalizar y ensalzar las publicaciones, más suaves, que sean más afines a ellos. Del mismo modo, pueden aprovechar para difundir propaganda redireccionando a los usuarios a otros lugares del ciberespacio (Weimann, 2018).

Asimismo, como se indicó en la reciente investigación de Esteve, Moneva y Miró-Llinares (2019) los usuarios de Twitter tienen la opción de interactuar de forma anónima, ocultando su identidad a través de pseudónimos o de nombre falsos (Peddinti et al., 2014), lo que les facilita expresar opiniones y realizar publicaciones sensibles sin miedo a ser identificados (Peddinti et al., 2017). Aunque no todos los usuarios de Twitter se ocultan tras un velo de anonimato para cometer conductas delictivas o desviadas, los trabajos de Peddinti y colaboradores (2017b; 2017a) muestran que existe una relación entre el anonimato de los usuarios y el contenido de carácter pornográfico, homófobo e islamófobo que se publica desde sus cuentas. Tanto el anonimato que proporciona esta red social como la facilidad con la que se accede a ella, y que permite que usuarios con diferentes características personales se sientan identificados con ciertas ideologías (Perry & Olsson, 2009), han favorecido que Twitter se convierta en una plataforma donde algunos usuarios emiten mensajes radicales y de odio que permanecen fijados a lo largo del tiempo, alcanzando a un público masivo, incrementando así su lesividad (Miró-Llinares et al., 2018).

A pesar de estas características homogéneas, Twitter es un ciberlugar heterogéneo donde las diferentes cosas suceden de distinta manera y en distintos microlugares. Por ejemplo, como indican Miró-Llinares y Rodriguez (2016), la distribución temporal de los mensajes no se produce de forma aleatoria, al igual que existen unos proxenetas con más seguidores que otros (Lara-Cabrera et al., 2017). Por lo tanto, la investigación se debe centrar en los microlugares digitales, como las cuentas de los usuarios y los tweets publicados, y su relación con el evento criminal para identificar los patrones ambientales del mismo. En este sentido, Twitter puede verificar con una insignia azul aquellas cuentas cuya información de registro coincide con la identidad de una personalidad pública. Del mismo modo, como ya hemos indicado, el usuario puede incluir una breve biografía en su perfil además de activar una opción para geolocalizar los tweets (Miró-Llinares et al., 2018). Igualmente, los usuarios pueden crear sus propias "listas", agrupando las cuentas de Twitter que hayan seleccionado por diferentes criterios (trabajo, viajes, noticias, etc.). Cada lista genera su propio flujo de publicaciones o timeline, que será diferente al general. El usuario, además de confeccionar diferentes listas, también puede ser seguidor de listas públicas que hayan sido creadas por otras personas e incluso controlar otras en las que se le haya incluido. El número de listas en las que se incluye una cuenta se indica en su perfil junto con otras características, como el número de seguidores y cuentas seguidas, así como el número de tweets publicados.

Asimismo, los mensajes que se publican en Twitter se definen por una variedad de elementos que lo configuran. Los tweets no deben superar los caracteres permitidos, independientemente que sean alfanuméricos o incluso emojis (pequeños iconos). La variedad de estos elementos, junto con otros como las menciones, los retweets, los hashtags, la inclusión de un hipervínculo, vídeos, imágenes, un GIF o un enlace a un

sitio externo, van a definir el contenido del microespacio de Twitter y su interacción entre los usuarios.

En definitiva, nos encontramos ante elementos situacionales que nos servirán para clasificar el contenido de un tweet (como neutro o violento), enlazar publicaciones y establecer tendencias comunicativas relacionadas con la comunicación violenta y el discurso de odio, para conocer los patrones ambientales que pueden identificar las características de este fenómeno con el objetivo de prevenir, controlar o disminuir sus efectos en la sociedad (Miró-Llinares et al., 2018).

# Parte III

# ESTUDIO EMPÍRICO

# Capítulo 5

# ESTUDIO EMPÍRICO

# ANÁLISIS DE LAS CARACTERÍSTICAS AMBIENTALES DE TWITTER PARA LA PREVENCIÓN DE LA COMUNICACIÓN VIOLENTA Y EL DISCURSO DE ODIO

"La ciencia siempre vale la pena porque sus descubrimientos, tarde o temprano, siempre se aplican"

(Severo Ochoa)

De acuerdo con el desarrollo teórico realizado en la primera parte de esta investigación, nos planteamos estimar la prevalencia de la comunicación violenta y el discurso de odio en la red social Twiiter, así como elaborar un modelo predictivo que permita diferenciar la comunicación neutral de la comunicación violenta y el discurso de odio a partir de las características ambientales de sus metadatos.

La investigación se realiza con una muestra de tweets recogidos después de tres atentados terroristas: El atentado a la revista Charlie Hebdo en París en 2015, el atentado en Bruselas en 2016 y el atentado en la ciudad de Londres en 2017. El interés fundamental del estudio es desarrollar una metodología que pueda discriminar los mensajes neutros de los mensajes violentos en Twitter, a través de la creación de un nuevo algoritmo, utilizando las variables ambientales de los tweets publicados. De este modo, conseguiremos una herramienta que pueda discriminar los mensajes neutrales de los violentos sin necesidad de analizar su contenido semántico. Así, se podrán

identificar los mensajes neutros conociendo cuáles son los micro-espacios más seguros en Twitter y se podrá reducir la muestra objeto de estudio por parte de los expertos (policía, proveedores de servicio, etc.) para proceder a su eliminación, ya sea total o parcial. Hecho que también nos será de utilizad para elaborar diferentes estrategias preventivas en este sentido.

En el presente capítulo se mostrarán los objetivos que se pretenden conseguir con el desarrollo de la investigación. También se indicarán las diferentes hipótesis planteadas y, posteriormente, se desarrollará la metodología utilizada. Por último, se presentarán los resultados obtenidos.

## 1. Objetivos e hipótesis

### *1.1. Objetivos*

La investigación tiene como objetivo principal el desarrollo de una metodología que permita diferenciar la comunicación neutral de la comunicación violenta y el discurso de odio. Para ello, se utilizarán los metadatos o "variables ambientales" como fuente válida para determinar si los mensajes de un tweet son neutrales o de odio. Los metadatos se definen como la información que describe a otros datos (McCarthy, 1982). En el caso de un tweet, los metadatos harían referencia a la información que lo describe. Es decir, a los datos relativos a la fecha de publicación, la cantidad de retweets, número de seguidores, geoposicionamiento, etc.

Otro de los objetivos que se pretende alcanzar, empleando esta metodología, es el de superar las limitaciones que va a asociadas a un idioma consiguiendo una precisión semejante a la de otros enfoques centrados en el análisis de los contenidos de los mensajes. Los trabajos que se han desarrollado con un enfoque de análisis semántico y

en la sintaxis que emplean búsqueda de palabras clave (Décary-Hétu & Morselli, 2011), bolsas de palabras (Waseem & Hovy, 2016) o análisis de sentimiento (Agarwal & Sureka, 2017) tan solo logran precisiones cercanas al 70%.

Para poder alcanzar estos objetivos principales, se presentan los siguientes específicos:

- Obtener datos primarios de Twitter dentro de un contexto que favorezca la publicación de mensajes con una elevada carga ideológica y emocional que potencie la interacción entre los usuarios y la publicación de mensajes violentos y discurso de odio.
- Categorizar los mensajes obtenidos a través del análisis humano basado en la taxonomía de la comunicación violenta y el discurso de odio (Miró-Llinares, 2016).
- Estimar la prevalencia de la comunicación violenta y el discurso de odio en Twitter.
- Realizar un análisis descriptivo de las variables independientes para conocer sus características entre los distintos tipos de mensajes.
- Identificar, a través de técnicas supervisadas de *machine learning*, los patrones de las variables ambientales (i.e., metadatos) que están presentes en los mensajes de comunicación violenta y discurso de odio y en los mensajes neutros para facilitar su detección automática.

### *1.2. Hipótesis*

Una vez realizado el planteamiento desarrollado en la primera parte de esta investigación, se formulan 3 hipótesis relacionadas con la comunicación violenta y el discurso de odio en Twitter, basadas principalmente en la idea de que los metadatos asociados a la interacción y la estructura de los tweets (variables ambientales) son especialmente relevantes para identificar su contenido.

$H_1$. La mayoría de la comunicación emitida a través de Twitter es de carácter neutral, por lo que la prevalencia de la comunicación violenta y el discurso de odio es escasa.

$H_2$. La comunicación neutral y la comunicación violenta y el discurso de odio poseen características ambientales distintas, lo que se refleja en una distribución diferencial de sus metadatos.

$H_3$. Las características ambientales obtenidas de los metadatos sirven para elaborar un modelo predictivo que permite diferenciar la comunicación neutral de la comunicación violenta y el discurso de odio.

## 2. Método

### *2.1. Muestra*

La muestra empleada para la realización de esta investigación está compuesta por un total de 453.924 tweets, divididos en tres muestras recogidas tras diferentes ataques terroristas (Tabla 5).

La primera muestra (*N*=229.181) se obtuvo entre los días 7 y 12 de enero de 2015, tras el ataque perpetrado, en París el 7 de enero, contra Charlie Hebdo, semanario satírico francés. Los hermanos de nacionalidad francesa, padres argelinos y miembros

de Al-Qaeda, Said y Chérif Kouachi, tirotearon al grito de "Alá es el más grande", la sede de la revista matando a doce personas e hiriendo a otras once. Posteriormente abatieron, con sus fusiles automáticos Kalashnikov, al policía Ahmed Merabet. Las Fuerzas y Cuerpos de Seguridad del país derribaron a los terroristas unos días después.

La segunda muestra (*N*=23.863) se obtuvo entre los días 22 y 28 de marzo, tras los atentados terroristas cometidos la mañana del 22 de marzo de 2016, en el aeropuerto internacional Zaventem y en la estación de metro Meelbeck de Bruselas. Los responsables del ataque en el aeropuerto fueron cuatro miembros de Dáesh: Najim Laachraoui, Ibrahim El Bakraoui (se inmolaron en el acto) y Mohamed Abrini. Éste último fue detenido porque su maleta no llegó a estallar, aunque lo tuvieron que dejar libre por falta de pruebas. El último integrante de la operación, Khalid El Bakraoui, fue quien se inmoló en el metro. Dejaron la lamentable cifra de 32 muertos y 300 heridos.

La tercera y última muestra (*N*=200.880) se obtuvo, entre los días 3 y 5 de junio de 2017, tras los ataques perpetrados, el 3 de junio, en el puente de Londres y en el Mercado Borough de la ciudad. Los terroristas miembros de Dáesh, Khurum Shazad Butt, Rachid Redouane y Yusef Zaghaba, mataron al menos a 7 personas y dejaron 48 heridos como consecuencia del atropello masivo realizado en la acera del puente de Londres y los apuñalamientos en el mercado. Fueron abatidos por la policía.

*Tabla 5. Conjunto de tweets de las muestras obtenidas tras los atentados de Charlie Hebdo, Bruselas y Londres*

| Atentado | Lugar | Año | Muestra |
|---|---|---|---|
| Charlie Hebdo | París | 2015 | 229.181 |
| Bruselas | Bruselas | 2016 | 23.863 |
| Londres | Londres | 2017 | 200.880 |

### *2.2. Procedimiento de selección de la muestra*

Los sistemas informáticos organizan las fuentes de datos de diferentes maneras. Los datos se pueden ordenar, por ejemplo, en una hoja de cálculo de excel o en un archivo de texto plano, es decir, únicamente caracteres sin formato. A los datos de las redes sociales, como Twitter, solo se puede acceder si los desarrolladores de ésta así lo permiten y estarán organizados de la forma en la que ellos lo hayan establecido. Las APIs son aplicaciones que contienen funciones para crear, tarea que se realiza por los desarrolladores para organizar y otorgar los datos que consideren, y para acceder a los datos, tarea llevada a cabo por los usuarios con permisos de desarrollador a través del uso de lenguajes de programación.

En este sentido, los datos para esta investigación se recogieron utilizando la API *Streaming* de Twitter. Esta API captura los datos que se están publicando en tiempo real, pudiendo almacenar hasta el 1% de los tweets publicados en función de los criterios de búsqueda preestablecidos. En nuestra investigación, se seleccionaron tres hashtags por evento para proporcionar un muestreo equilibrado (véase Miró-Llinares, 2016), que atienden a tres criterios: relacionado con el evento, con las expresiones solidarios y con las islamistas (Tabla 6).

- Los hashtags relacionados con el evento de forma neutral: #CharlieHebdo, #Bruselas y #London.
- Los hashtags para el contenido de solidaridad: #JeSuisCharlie, #PrayForBruselas y #PrayForLondon.
- Los hashtags representativos de las expresiones radicales, concretamente de islamofobia: #stopIslam.

Para la obtención de la muestra, también se seleccionaron estas palabras clave, es decir, esos 6 hashtags, pero eliminando el propio hashtag para obtener mayor número de tweets.

*Tabla 6. Hashtags empleados para filtrar la captura de datos*

| Evento | Hashtag evento | Hashtag solidario | Hashtag radical |
|---|---|---|---|
| Charlie Hebdo | #CharlieHebdo | #JeSuisCharlie | #stopIslam |
| Bruselas | #Bruselas | #PrayForBruselas | #stopIslam |
| Londres | #London | #PrayForLondon | #stopIslam |

Los eventos seleccionados para el estudio son los relativos a los atentados terroristas perpetrados en la revista satírica de Charlie Hebdo en 2015, en el aeropuerto y en el metro de Bruselas en 2016 y en el puente y mercado de Londres en 2017. Estos tres eventos son apropiados porque, debido al elevado grado de sensibilidad que creó entre los usuarios, provocó la difusión de miles de mensajes en Twitter. Además, la diferencia temporal entre los eventos nos permite comprobar la aplicabilidad genérica de la metodología propuesta. Respecto al idioma de los tweets recogidos, los de la muestra de Charlie Hebdo y Bruselas son mensajes en español y los últimos de Londres, es inglés.

Así pues, se obtuvo un total de 453.924 tweets con sus correspondientes metadatos descritos en la siguiente tabla (Tabla 7).

*Tabla 7. Metadatos de un tweet*

| Campos | Tipo Campo | Valores | Definición |
|---|---|---|---|
| Text | String | 140 caracteres | Contenido del *tweet*. |
| Retweet_count | Int | 0...n | Número de veces que el *tweet* ha sido *retweeteado*. |
| Favorited | Boolean | True/False | Indica si el *tweet* le ha gustado al usuario autentificado. |
| Truncated | Boolean | True/False | Indica si el campo "*Text*" ha sido truncado.<br>El campo "*Text*" no puede exceder los 140 caracteres, |

| Campos | Tipo Campo | Valores | Definición |
|---|---|---|---|
| | | | a excepción de que incluya enlaces o imágenes. Si el texto es truncado terminará en puntos suspensivos. |
| Id_str | String | Id del creador del *tweet* | Id del usuario creador del *tweet*. |
| In_reply_to_screen_name | String | NA/Nombre usuario del *tweet* original | Si el *tweet* es una respuesta, este campo contiene el nombre de usuario del *tweet* original. |
| Source | String | HTML | Cadena en formato HTML. |
| Retweeted | Boolean | True/False | Indica si el *tweet* ha sido *retweeteado* por el usuario autentificado. |
| Created_at | String | Hora UTC | Hora UTC cuando el *tweet* ha sido creado. |
| In_reply_to_status_id_str | String | NA/Id del *tweet* original | Si el *tweet* es una respuesta este campo contiene el Id del *tweet* original. |
| In_reply_to_user_id_str | String | NA/Id del usuario original | Si este *tweet* es una respuesta este campo contiene el Id del usuario original. |
| Lang | String | UND/código abreviado del idioma | Idioma del *tweet*. |
| Listed_count | Int | 0...n | Número de listas públicas en las que el usuario es miembro |
| Verified | Boolean | True/False | Si es "*True*" indica que la cuenta ha sido verificada. |
| Location | String | Cadena de texto | Ubicación definida por el usuario para el perfil de esa cuenta. No es necesariamente un lugar ni analizable. |
| User_id_str | String | Cadena de texto | Id del usuario. |
| Description | String | Cadena de texto | Cadena de texto definida por el usuario para describir su cuenta. |

| Campos | Tipo Campo | Valores | Definición |
|---|---|---|---|
| Geo_enabled | Boolean | True/False | Indica si el geoposicionamiento del *tweet* está activado o, por el contrario, está desactivado. |
| User_created_at | String | Hora UTC | Hora UTC cuando el usuario fue creado. |
| Statuses_count | Int | 0...n | Número de *tweets* (incluyendo *retweets*) emitidos por el usuario. |
| Followers_count | Int | 0...n | Número de seguidores de la cuenta. |
| Favourites_count | Int | 0...n | Número de tweets que este usuario ha marcado como favorito en la vida de la cuenta. |
| Protected | Boolean | True/False | Cuando es "*True*" indica que el usuario ha decidido proteger sus *tweets*. |
| User_url | String | Url | Url del usuario. |
| Name | String | 20 caracteres | Nombre del usuario. |
| Time_zone | String | Cadena de texto | Zona horaria que ha declarado el usuario. |
| User_lang | String | Cadena de texto | Idioma que ha declarado el usuario. |
| Utc_offset | Int | 0…n | Desplazamiento desde GMT/UTC en cuestión de segundos. |
| Friends_count | Int | 0...n | Número de usuarios que esta cuenta está siguiendo. |
| Screen_name | String | 15 caracteres | Alias del usuario. |
| Country_code | String | Código del país | Código del país acortada. |
| Country | String | Cadena de texto | Nombre del país. |
| Place_type | String | Country/City/... | Tipo de ubicación. |
| Full_name | String | Cadena de texto | Nombre del lugar. |
| Place_name | String | Cadena de texto | Tipo de lugar (país, vecindario, calle, …) |
| Place_id | Int | 0…n | Identificador del lugar. |
| Place_lat | Float | - | Latitud del lugar. |

| Campos | Tipo Campo | Valores | Definición |
|---|---|---|---|
| Place_lon | Float | - | Longitud del lugar. |
| Lat | Float | - | Latitud del *tweet* creado. |
| Lon | Float | - | Longitud del *tweet* creado. |
| Expended_url | String | Url | Url multimedia completa. |
| Url | String | Url | Url que representa la ubicación del lugar. |

### *2.3. Preprocesamiento de los datos*

Prácticamente todos los procesos de preparación de datos consisten en algún tipo de transformación (Hernández et al., 2004). Algunos de ellos de mayor complejidad que otros, pero siempre especialmente importantes. Existen muchos estudios sobre las técnicas de preprocesamiento porque, esta fase, es muy diferente en cada problema que se aborda (Wasilewska & Menasalvas, 2008). Algunas de las pautas más frecuentes en el tratamiento de los datos de una muestra se enfocan en los metadatos, que serían las columnas y en los casos, las filas. En el primer caso, un tratamiento clásico es la agregación de características que consiste en crear nuevos campos para mejorar la calidad, visualización o compresibilidad del conocimiento extraído. También es frecuente tener que recurrir a la discretización de variables numéricas o de tipo fecha, para establecer rangos o segmentos que faciliten la ejecución de algunos modelos analíticos, incapaces de trabajar con variables numéricas continuas. También se suelen llevar a cabo operaciones como la eliminación de variables porque acumulan una dispersión de valores intratables en la práctica como, por ejemplo, el cuerpo del tweet. O por ser claves primarias o candidatas que identifican unívocamente a los registros como, por ejemplo, el nombre de usuario. Aquellas variables, que fruto de algún tipo de selección o sesgo presenten un único valor de forma muy mayoritaria, por encima del

90% de la muestra, también suelen ser eliminados por ser considerados, a efectos prácticos, variables unievaluadas. Un caso particular lo presentan los conjuntos de datos con variables con un elevado porcentaje de nulos. Normalmente, si no tiene sentido tratar el nulo como un valor más de la variable o si la concentración de nulos supera el 25-30%, dicha variable también suele ser eliminada. En casos especiales, donde las variables atienden a algún tipo de gradiente o distribución conocida, los nulos pueden ser reemplazados por interpolación. Por ejemplo, valores ausentes en una serie de temperaturas.  En el segundo caso, el enfoque del tratamiento de las filas, una técnica es la eliminación de aquellos casos que presenten nulos o valores claramente erróneos en variables potencialmente críticas para el análisis. Por ejemplo, fecha de creación de un usuario posterior a la fecha actual. También se suele utilizar la eliminación de aquellos casos que presentan un alto porcentaje de nulos (por encima del 25-30%). Por otro lado, los valores fuera de rango (*outliers*) suelen ser tratados como valores nulos y se sigue la estrategia de ser reemplazarlos cuando sea posible o conveniente o eliminando su correspondiente fila.

En nuestra investigación, las acciones de preprocesamiento llevadas a cabo sobre los metadatos descritos (Tabla 7) han sido las siguientes:

- Eliminación de las siguientes variables por su alta dispersión de valores y campos de texto: source, Name, Screen_name, description y Agent.
- Eliminación de variables identificativas del usuario y del tweet: IdUser.
- Eliminación de las variables con un alto valor de nulos (superior al 70%): location, url, Time_zone, Geo_Enabled, Coordinate, Entities, Urls, In_reply_to_screen_name, Geo_Lon y Geo_Lat.

- Eliminación de las siguientes variables, tras ser empleadas para la creación de variables derivadas de mayor interés analítico: text, create_twt y create_user.

Otra de las acciones realizadas es la extracción de nuevas variables ambientales a partir de otras. Estas nuevas variables son:

- Texto del tweet (*text*):
- *Mention_count* (numérica): Número de menciones incluidas en el texto del tweet.
- *Hashtag_count* (numérica): Número de hashtags utilizados.
- *Link* (binario): Indica si en el texto se incluye una URL o no.
- *Text_count* (numérica): Número de caracteres empleados para escribir el mensaje.
- Creación del tweet (*created_at*):
- *Minute_count* (numérica): Minutos transcurridos entre el evento y la publicación del tweet.
- Creación del usuario *(user_created_at*):
- *Day_count* (numérico): Días transcurridos entre la creación del usuario y el último tweet recogido en la muestra.

Así pues, las variables ambientales que finalmente se han seleccionado para el estudio son: *mention_count, hashtag_count, link, text_count, minute_count, description, geoenabled, day_count, listed_count, statuses_count, followers_count, friends_count, favourites_count y retweet_count*

### *2.4. Variables*

#### *2.4.1 Variable Dependiente*

El objetivo de esta investigación es la detección y prevención de la comunicación violenta y el discurso de odio, por lo tanto, ésta será nuestra variable dependiente. Para ello, se realizó una lectura minuciosa de cada uno de los tweets de la muestra, con la siguiente dicotomía (0=mensaje neutro; 1=mensaje odio) para determinar si el contenido era neutro o una comunicación violenta. Se eligió este método en lugar de los enfoques semánticos o sintácticos (ej., Bolsa de Palabras) porque, como ya se ha indicado, éstos han mostrado debilidades al tratar mensajes específicos como el humor o la ironía (Farias-Hernandez et al., 2016; Reyes et al., 2013).

Como se indicó en la primera parte de la investigación, este estudio plantea la clasificación del contenido de la muestra de los tweets, sobre la base de un concepto amplio de expresiones de odio. En ella se recogen todas las expresiones consideradas, según la taxonomía elaborada por Miró-Llinares (2016), como comunicación violenta o de odio. A partir de esta clasificación, para que el contenido de un tweet sea considerado como discurso de odio, debe incluir las siguientes categorías:

1. Incitación directa/amenaza de violencia
2. Glorificación de la violencia física
3. Ataque al honor y a la dignidad humana
4. Incitación a la discriminación/odio
5. Ofensa a la sensibilidad colectiva.

Para ello, se optó por la clasificación de cada una de las muestras a través de una interpretación subjetiva del texto siguiendo una metodología de validación interjueces,

asumiendo las limitaciones derivadas de este método. Los responsables encargados de la codificación de los tweets fueron investigadores seleccionados previo entrenamiento sobre una muestra de 100 tweets, de cada una de las bases de datos (Charlie Hebdo, Bruselas y Londres). Posteriormente, se realizó una prueba piloto de 200 tweets para comprobar la fiabilidad en la valoración de los jueces. Para subsanar el efecto del análisis subjetivo de los mensajes por parte de los jueces y garantizar la conformidad en las evaluaciones, se aplicó el coeficiente Kappa (Cohen, 1960), encargado de medir el grado de acuerdo. Cuando el resultado del valor de K es <0,20 estamos ante una concordancia pobre; sería débil entre el 0,21 y 0,40; moderada entre 0,41 y 0,60; de 0,60 a 0,80 buena y de 0,81 a 1 muy buena (Landis y Koch, 1977). De acuerdo con los criterios establecidos, se obtuvo un acuerdo muy satisfactorio entre los pares de jueces (Tabla 8).

*Tabla 8. Resultados de la aplicación del coeficiente Kappa de los tres pares de jueces en las tres muestras obtenidas*

| Charlie Hebdo | | Bruselas | | Londres | |
|---|---|---|---|---|---|
| Grupo | *k* | Grupo | *k* | Grupo | *k* |
| Grupo 1 | 0.98 | Grupo 1 | 0.92 | Grupo 1 | 0.81 |
| Grupo 2 | 0.86 | Grupo 2 | 0.90 | Grupo 2 | 0.89 |
| Grupo 3 | 0.98 | Grupo 3 | 0.89 | Grupo 3 | 0.88 |

En las tres muestras se han eliminado los tweets y retweets para proteger los mensajes originales y obviar las réplicas que pudieran ser redundantes (Esteve et al., 2018; Miro-Llinares & Rodriguez-Sala, 2016; Miró Llinares, 2016). De este modo, se obtuvieron 2.274 tweets originales de la muestra de Charlie Hebdo, 1.935 de Bruselas y 35.433 tweets en la muestra de Londres listos para codificar. Después de que los jueces clasificaran estos mensajes, los duplicados se volvieron a doblar en el conjunto de datos para calcular la prevalencia del discurso de odio en las muestras de tweets. El resultado

fue: 1.966 (0,86%) de 229,181 en Charlie Hebdo; 4,469 (18,7%) de 23.863 y 9488 (4,7%) de 200.880 en Londres (Tabla 9).

*Tabla 9. Conjunto de tweets de la muestra de los atentados de Charlie Hebdo, Bruselas y Londres*

| Variable dependiente | Muestra de tweets | | | | | |
|---|---|---|---|---|---|---|
| | Charlie Hebdo | % | Bruselas | % | Londres | % |
| Cv_Do | 1.966 | 0,86 | 4.473 | 18,72 | 9.488 | 4,74 |
| Neutral | 227.215 | 99,14 | 19.390 | 81,28 | 191.392 | 92,26 |
| TOTAL | 229.181 | 100 | 23.863 | 100 | 200.880 | 100 |

### *2.4.2 Variables Independientes*

Las variables independientes están construidas a partir del preprocesamiento de los datos y la eliminación de las variables de dispersión y las variables con porcentajes de nulos superiores a 25-30% (Hernández et al., 2004).

Por un lado, se han seleccionado aquellas variables que están relacionadas con el anonimato y la visibilidad de las cuentas y, por el otro, con la estructura y la interacción de los tweets. De este proceso, se han obtenido las 14 variables independientes que se utilizarán en esta investigación (Tabla 10).

Anonimato. En la categoría de anonimato, se han incluido las variables que proporcionan información sobre la persona que está detrás de esa cuenta, como sus intereses, trabajo, hobbies o lugar de residencia:

- Description: Esta variable hace referencia al texto que escribe el usuario para describir su cuenta (biografía). Se ha modificado porque la API recoge el texto completo y el análisis hubiera implicado una interpretación subjetiva. Por ello, se aplicó una dicotomía (0=no tiene;1=tiene biografía).

- Geo_enable**:** Esta variable hace referencia a la activación del geoposicionamiento del tweet. Es binaria, por lo tanto, se codificará si está activado o no (0=no está activado; 1=sí está activado)

Visibilidad. En la categoría visibilidad se ha incluido un segundo conjunto de variables que miden la visibilidad de actividades de los usuarios en Twitter, como el período de actividad del usuario en la red social, la publicación de mensajes y las diferentes formas de interacción con otros usuarios (Miró-Llinares et al., 2018).

- Day_count: Esta variable independiente es nueva. Es decir, se ha obtenido a partir de otra variable (user_created_at) extraída directamente de la muestra de la API de Twitter. Es numérica y hace referencia a los días transcurridos entre la creación del usuario y el último tweet reconocido en la muestra.
- Listed_count: Es una variable numérica. Hace referencia al número de listas públicas en las que el usuario es miembro.
- Statuses_count**:** Es una variable numérica. Indica el número de tweets (incluidos rtw) que ha enviado el usuario.
- Followers_count: Es una variable numérica y hace referencia al número de seguidores que tiene el usuario (followers).
- Friends_count**:** Es una variable numérica y hace referencia al número de usuarios que está siguiendo esa cuenta. Es decir, los usuarios a los que sigue (followings).
- Favourites_count: Es una variable numérica y hace referencia al número de tweets que este usuario ha marcado como favoritos en la vida de su cuenta.

Los propios tweets y sus metadatos asociados también se han identificado como posible predictores de la difusión del discurso de odio. Algunos de estos elementos

están relacionados con la interacción que genera un tweet, mientras que otros determinan su estructura.

Interacción. Dentro de la categoría de interacción, se han incluido algunos elementos interactivos que favorecen la participación de los usuarios en las actividades de difusión, junto con el momento de la publicación del tweet.

- Mention_Count: Esta variable independiente es nueva. Es decir, se ha obtenido a partir de otra variable (*text*), extraída directamente de la muestra de la API de Twitter. Es numérica y hace referencia al número de menciones (a través de la @) que están incluidas en el texto del tweet.
- Hashtag_Count: Esta variable independiente es nueva. Se ha obtenido a partir de otra variable (*text*), extraída directamente de la muestra de la API de Twitter. Es numérica y hace referencia al número de hashtag (conjunto de caracteres precedidos por una almohadilla #) que se han utilizado en el tweet.
- Link: Esta variable independiente es nueva. Es decir, se ha obtenido a partir de otra variable (*text*), extraída directamente de la muestra de la API de Twitter. Es binaria y hace referencia a la existencia de enlaces externos en el tweet (0=no hay enlace externo; 1=sí hay enlace externo).
- Retweet_count: Esta variable independiente se obtiene directamente de la información de la API. Es una variable numérica y hace referencia al número de veces que este tweet ha sido retweeteado.
- Minute_Count: Esta variable independiente es nueva. Es decir, se ha obtenido a partir de otra variable (*created_at*) extraída directamente de la muestra de la API de Twitter. Es numérica y hace referencia a los minutos transcurridos entre el evento y la publicación del tweet.

Estructura. La categoría de estructura comprende la variable que limitan la longitud del texto y, por consiguiente, del contenido del mensaje.

- Text_Count**:** Esta variable independiente es nueva. Es decir, se ha obtenido a partir de otra variable (text), extraída directamente de la muestra de la API de Twitter. Es numérica y hace referencia al número de caracteres empleados para escribir el mensaje. Para ello, se elaboraron breves guiones para identificar tanto la codificación de los emojis como las cadenas de caracteres que componen los enlaces externos y, posteriormente, extraerlos del cuerpo del mensaje. De este modo, es posible llevar a cabo un conteo de caracteres para determinar la longitud real de un mensaje.

*Tabla 10. Variables ambientales de los tweets divididas en las categorías anonimato, visibilidad, interacción y estructura.*

| Id | Nombre | Tipo | Descripción |
|---|---|---|---|
| | Anonimato | | |
| 1 | Description | Dicotómica | Descripción en la bibliografía en el perfil de la cuenta (0=no;1=sí) |
| 2 | Geo_enable | Dicotómica | La geolocalización del tweet está actividad (0=no;1=sí) |
| | Visibilidad | | |
| 3 | Day_count | Numérica | Número de días desde que se creó la cuenta |
| 4 | Listed_count | Numérica | Número de listas públicas en las que el usuario es miembro |
| 5 | Statuses_count | Numérica | Número de tweets (incluidos retweets) publicados por el usuario |
| 6 | Followers_count | Numérica | Número de seguidores que tiene el usuario (followers) |
| 7 | Friends_count | Numérica | Número de usuarios que la cuenta está siguiendo (followings) |
| 8 | Favourites_count | Numérica | Número de tweets que este usuario ha marcado como favoritos en la vida de su cuenta |
| | Interacción | | |
| 9 | Mention_count | Numérica | Número de menciones incluidas en el texto del tweet (a través de la @) |
| 10 | Hashtag_count | Numérica | Número de hashtags incluidos en el texto del tweet |
| 11 | Link | Dicotómica | Enlaces externos en el tweet (0=no; 1=sí) |
| 12 | Retweet_count | Numérica | Número de veces que este tweet ha sido retweeteado |
| 13 | Minute_count | Numérica | Número de minutos que han pasado desde que ocurrió el evento y se publicó el tweet |
| | Estructura | | |
| 14 | Text_count | Numércia | Número de caracteres en el mensaje (excluyendo enlaces externos, emoji y retweets |

### *2.5. Instrumento*

Una vez que se ha definido las variables, se presentan los instrumentos que se han empleado para el desarrollo del algoritmo.

El algoritmo de esta investigación se ha desarrollado en el lenguaje de programación Python. Este lenguaje de programación de código abierto está ganando una gran popularidad entre los científicos y desarrolladores de software (Robinson, 2017). A diferencia del lenguaje de programación R, que está destinado principalmente a la ciencia de datos, Python aparece con una gama más amplia de aplicaciones, como el desarrollo de páginas web, el acceso a bases de datos, desarrollo de softwares e incluso juegos (Hao & Ho, 2019). Python no es un lenguaje complicado, lo que significa que no recoge el código en binario. De este modo, el intérprete de Python traduce el guión a binario durante su ejecución en tiempo real. Este lenguaje de programación viene con funcionalidades básicas, pero, es cierto, que depende de paquetes externos para realizar casi todos los cálculos numéricos.

En este momento, quizás la interfaz más fácil de utilizar para programar con Python es el cuaderno Jupyter, que proporciona una interfaz interactiva para el intérprete de Python, siendo muy adecuada para la mayoría de los trabajos de análisis de datos. La forma más sencilla de obtener la herramienta Python, y sus paquetes centrales como Jupyter, es instalarlos a través de la página de Anaconda.

La distribución Anaconda es una de las plataformas más importantes del mundo relacionada con la ciencia de datos. Anaconda incorpora librerías, paquetes preinstalados y de código abierto que se pueden instalar en su repositorio. El usuario puede construir sus propios paquetes personalizados y compartirlos usando la nube Anaconda, PyPi y otros repositorios (Kadiyala & Kumar, 2017). Por lo tanto, hay

multitud de paquetes disponibles para su uso desde la nube, desarrollados por programadores de todo el mundo.

De todas las librerías de Anaconda, se ha utilizado Scikit-learn, el paquete de machine learning más completo. Aprovechando la amplia gama de aplicaciones de Python, Scikit-learn se convierte en un paquete cada vez más popular para las aplicaciones relacionadas con el machine learning (Hao & Ho, 2019). Además de ser de código abierto, tiene muchas características que lo hacen destacar entre otros programas similares. Existe un procedimiento de revisión comunitaria para identificar y decidir qué métodos de machine learning se deben incluir en el paquete. Este mecanismo es el que garantiza un equilibrio entre la amplia cobertura y la selección de los métodos de machine learning que contiene el paquete. Por otro lado, la aplicación del algoritmo de machine learning en Scikit-learn está optimizada para la eficiencia de la computación. Por último, Scikit-learn tiene un fuerte apoyo de la comunidad científica para la documentación, el seguimiento de errores y la garantía de calidad.

Respecto al Entorno de desarrollo (IDE) que se ha utilizado nos hemos decidido por Spyder (Scientific Python Develoment EnviRonment). Spyder es la aplicación dentro de Anaconda que proporciona un entorno de desarrollo científico de Python. Este programa facilita la edición avanzada, las pruebas interactivas, la depuración y las características de introspección en nuestra investigación, permitiéndonos escribir el código en Python para ejecutarlo posteriormente.

### *2.6. Procedimiento*

Para obtener la muestra de los tweets se realizó una captura de los mismos a través de la API de Twitter, tras tres atentados terroristas: el atentado a la revista satírica Charlie Hebdo en París en 2015, el atentado en Bruselas en 2016 y el atentado en la

ciudad de Londres en 2017. A continuación, siguiendo los criterios formulados en Miró-Llinares (2016), se procedió a categorizar los mensajes obtenidos a través del análisis humano. Además, se aplicó el coeficiente Kappa (Cohen, 1960), encargado de medir el acuerdo interjueces, para solventar el efecto del análisis subjetivo de los tweets por parte de los investigadores y salvaguardar la conformidad en las valoraciones.
Posteriormente, se preprocesaron los datos y se eliminaron las variables de dispersión y aquellas que tenían porcentajes nulos superiores a 25-30% (Hernández et al., 2004). De entre ellas, se seleccionaron 14 variables relacionadas, por un lado, con el anonimato y la visibilidad de las cuentas y, por el otro, con la estructura y la interacción de los tweets. De este manera, podemos analizar las reacciones de ciertos ciberlugares en Twitter y si existen diferencias entre unos y otros (Miró-Llinares & Johnson, 2018)

Para confirmar la relevancia de los lugares en el ciberespacio, se utilizó la técnica de clasificación Random Forest (Breiman, 2001), implementando un algoritmo que creó una serie de clasificadores para tweets que dividen la muestra en base a filtros generados por cada una de las variables incluidas en el modelo (es decir, los nodos). Estos clasificadores crecen a partir de un conjunto de datos aleatorios extraídos de la muestra principal para entrenar el modelo y sus parámetros. Uno de los problemas de estas técnicas es que en el momento de "aprender" de los comportamientos de una muestra de entrenamiento, puede "sobre aprender" de la variable dependiente mayoritaria. En nuestro caso, la variable dependiente mayoritaria de nuestra muestra es el elevado número de tweets que clasificaron como neutrales. Por ello, se realizó un balanceo de datos, una vez divida la muestra en muestra de entrenamiento y muestra de validación, al 50% a favor de la clase minoritaria. Es decir, de los tweets clasificados como comunicación violenta o discurso de odio. El balanceo al 50% solo es en la

muestra de entrenamiento compuesta por las muestras de Charlie Hebdo y Bruselas, ya que la muestra de validación es la base de datos completa de Londres (Tabla 11).

*Tabla 11.Muestras de entrenamiento y de validación*

| Clase | Muestra de entrenamiento | Muestra de validación |
|---|---|---|
| Neutral | 6.435 | 191.392 |
| Cv_Do | 6.435 | 9.488 |
| Total | 12.870 | 200.880 |

Este proceso de entrenamiento y prueba permite controlar los nodos anómalos o menos consistentes y, por lo tanto, el crecimiento de un árbol podado no desbordado. Para definir los parámetros más apropiados para nuestro algoritmo, se llevaron a cabo una serie de experimentos computacionales. Estos parámetros se ajustaron para reducir la sensibilidad del bosque a su valor (Túffery, 2011).

Al pasar por cada nodo, el modelo pregunta a cada clasificador si la muestra cumple con la condición establecida en él, con lo que se filtra la muestra principal y se crean dos submuestras: una que cumple con la condición y otra que no. El modelo selecciona el mejor filtro, entre todos los árboles, y promedia sus estimaciones individuales para generar la producción final (Miró-Llinares et al., 2018). De este modo, la técnica del Random Forest produce predicciones robustas, creando varios árboles de decisión que aprenden de un conjunto de entrenamiento predeterminado. Cuando la condición que define un nodo alcanza la máxima eficacia de clasificación, significa que el modelo ha llegado a un nodo de hoja, y clasifica la submuestra correspondiente a la misma clase: comunicación violenta y discurso de odio o contenido neutro. Esta técnica pretende demostrar que las variables ambientales del ciberlugar pueden ser utilizadas para clasificar de manera adecuada una parte de la muestra, contribuyendo a la automatización del proceso. Además, para evitar que los resultados sean influenciados por la composición del conjunto de entrenamiento, de manera positiva o negativa, se ha

procedido a la validación de los datos, a través de *K-fold cross validation*. A continuación, se presenta un esquema general de la metodología empleada en la investigación (Figura 9).

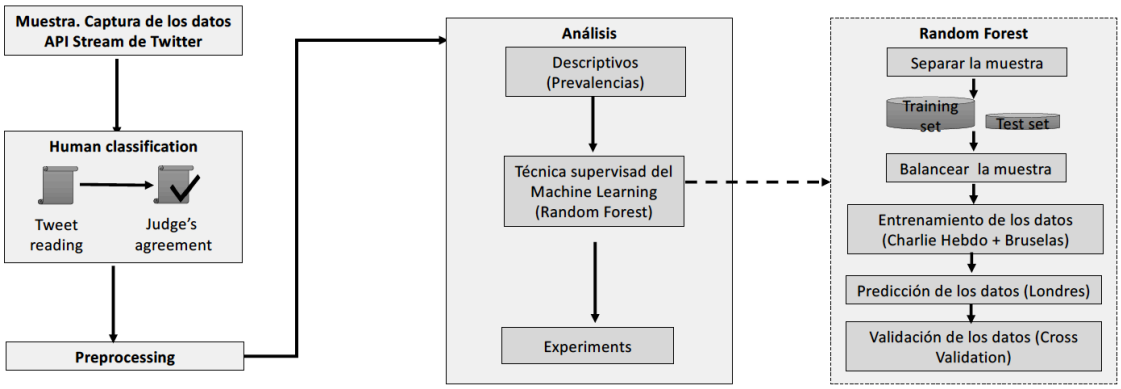


*Figura 9. Esquema de la metodología empleada en la investigación*

## 3. Resultados

El objetivo principal de nuestra investigación es la creación de un modelo predictivo basado en metadatos que permita diferenciar la comunicación neutral de la comunicación violenta y el discurso de odio.

No obstante, para el adecuado desarrollo del algoritmo es recomendable comenzar con el análisis descriptivo de la muestra de los mensajes utilizados en el estudio, así como de las variables ambientales.

### *3.1. Análisis descriptivo de las muestras obtenidas tras los atentados de Charlie Hebdo, Bruselas y Londres.*

En este apartado se ha realizado un análisis descriptivo de las muestras obtenidas tras los atentos de Charlie Hebdo, Bruselas y Londres a través del conteo de palabras, análisis de bigramas y el análisis de los principales hashstags.

### *3.1.1 Análisis semánticos*

En este apartado, se ha realizado un conteo de palabras, sin hashtags, de las tres muestras para conocer cuáles eran las que más se repetían.

En la gráfica de la muestra obtenida tras el atentado de Charlie Hebdo en 2015 se muestra las palabras que se repiten más de 5000 veces, siendo las más frecuentes: "libertad" con más de 30.000 menciones, seguido de "París" con unas 27.000 y "atentado" con un poco más de 25.000 publicaciones. En cuarta posición se encuentra la palabra "expresión" con 21.000 menciones, seguido de "Francia" y "hoy" (Figura 10). Como se puede observar, las palabras más utilizadas están relacionadas con el evento y la solidaridad.

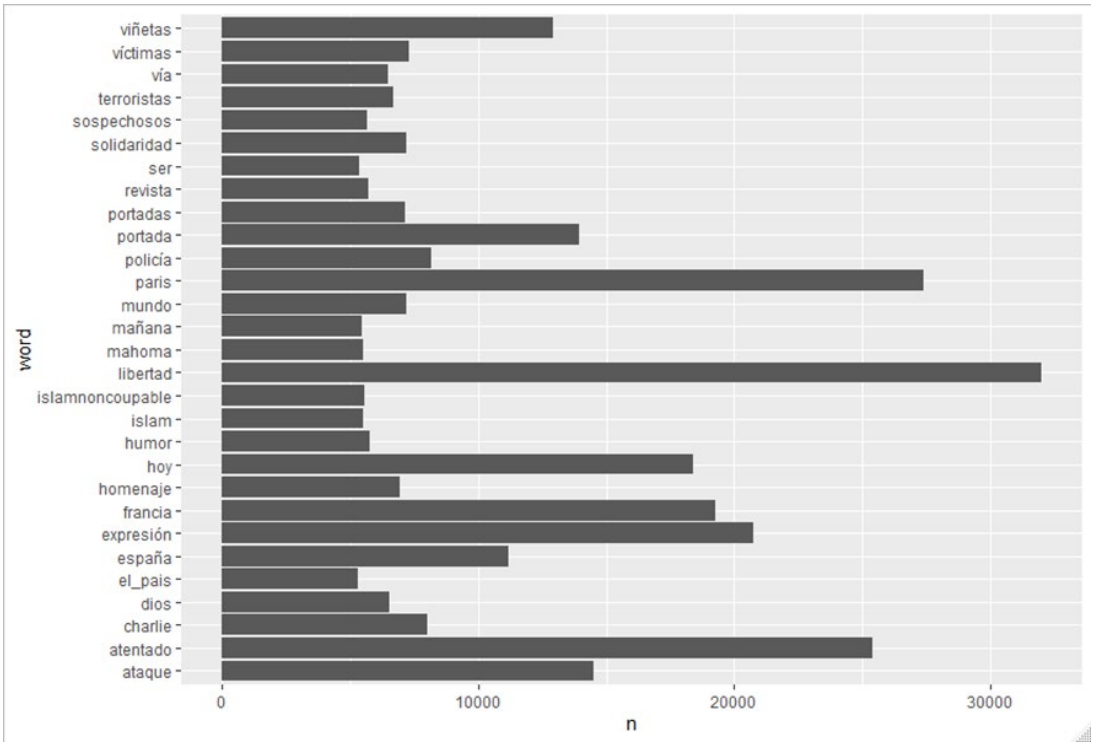


*Figura 10. Palabras que se repiten más de 5000 veces en los tweets recogidos de la muestra de Charlie Hebdo (2015)*

Así, en la gráfica de la muestra obtenida tras el atentado de Bruselas en 2016 se muestran las palabras que se repiten más de 500 veces, siendo las más frecuentes las relacionadas con el evento, pero también con la religión: "atentadobruselas" con más de

2.800 menciones, “atentados” que casi alcanza las 2.000 y, por último, “Jesuisbruxelles” que casi cuenta con 1.900 menciones. Le siguen palabras como “Bélgica”, “mundo”, “terrorismo”, “religión”, “terroristas”, “musulmanes” e “islamofobia” (Figura 11).

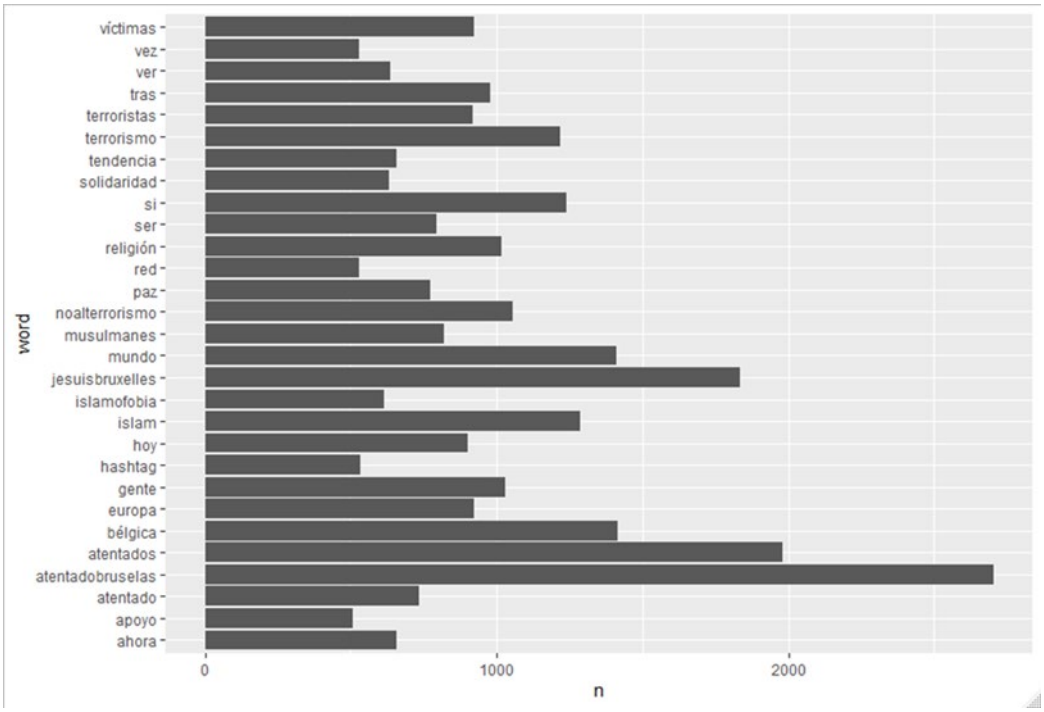


*Figura 11. Palabras que se repiten más de 500 veces en los tweets recogidos de la muestra de Bruselas (2016)*

En la siguiente figura se observa este conteo representado en una nube de palabras, en la que visualizamos, en mayor tamaño, las que más se repiten (Figura 12).


*Figura 12. Nube de palabras muestra Bruselas*

Del mismo modo, en la gráfica de la muestra obtenida tras el atentado de Londres en 2017, se indican las palabras que se repiten más de 5000 veces, siendo las más frecuentes: “Boroughmarket” con más de 34.000 menciones, seguido de “people” con más de 26.000 y “pólice” con 24.000 publicaciones, “attack” con unas 22.000 menciones y “god”, con 17.000. Le siguen palabras como “spilling”, “bless”, “brits” y “hmannella” (Figura 13).

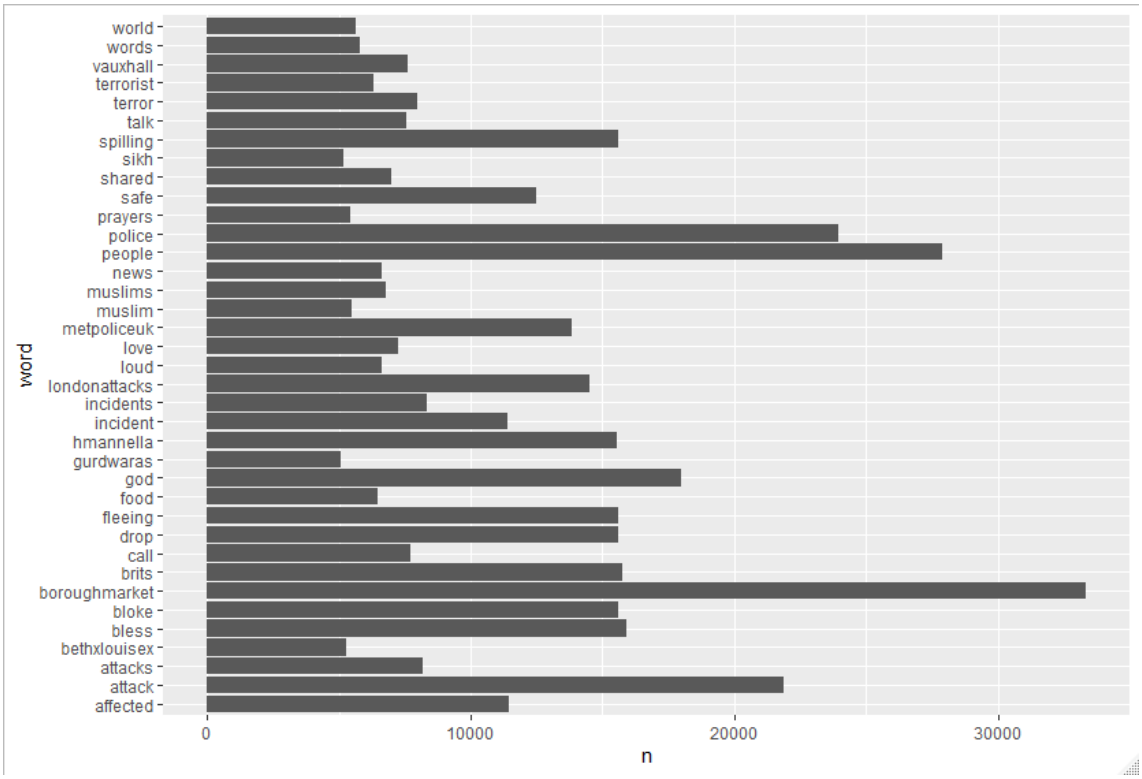


*Figura 13. Palabras que se repiten más de 5000 veces en los tweets recogidos de la muestra de Londres (2017)*

En la siguiente figura se observa este conteo representado en una nube de palabras, en la que visualizamos en mayor tamaño las que más se repiten (Figura 14).


*Figura 14. Nube de palabras muestra Londres*

Por otro lado, se ha realizado un análisis TF-IDF de las tres muestras para saber cuál es la frecuencia relativa de las palabras obtenidas en función del total, disminuyendo el peso de las palabras que no son importantes o se repiten con frecuencia, por ejemplo “la”, “y”, “de”. Como se puede observar las palabras que más peso tienen sobre el resto son “prayforbelgium”, “jesuischarlie” y “londonbridge”. En la muestra de Charlie Hebdo, se observa el peso de las palabras relativas al evento como “atentado”, “París”, seguidas de palabras relacionadas con la libertad de expresión. En la muestra relacionada con el atentado de Bruselas, se observa el peso de las palabras relacionadas con el evento como “atentadobruselas”, “Bruselas”; la solidaridad como “jesuisbruxelles”, “contra” y la religión. Por último, en la muestra de Londres, queda patente el peso de las palabras relacionadas con el sentimiento, como “prayforlondon”, “bless”, “brits” y “drop”

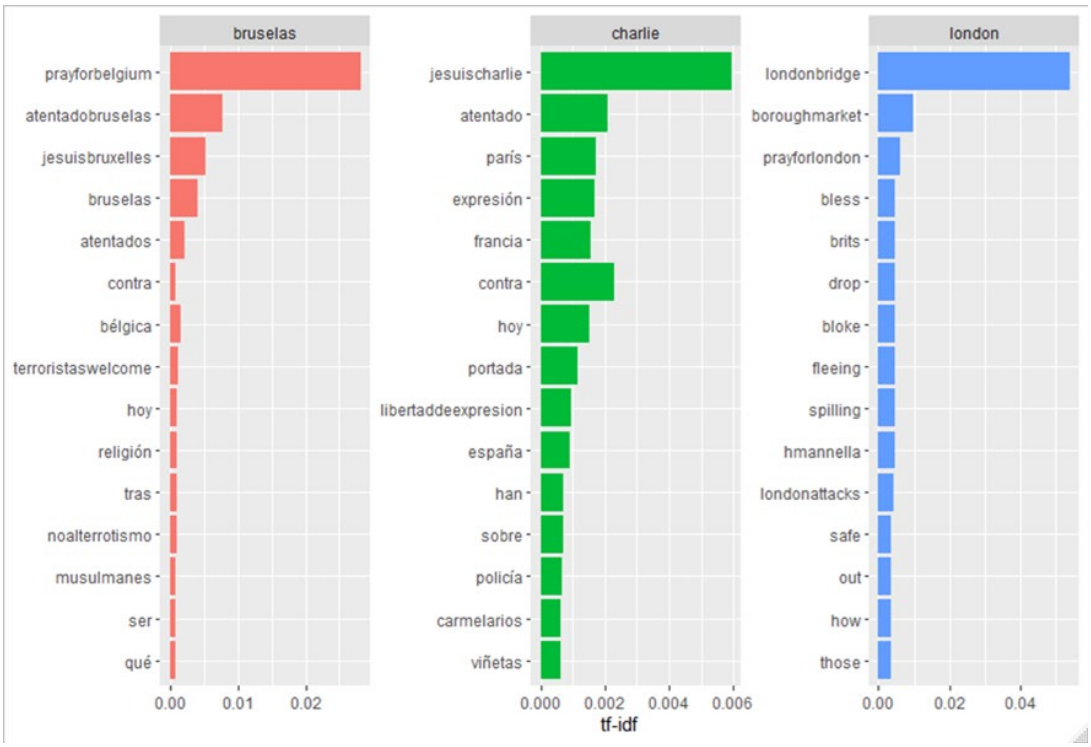


*Figura 15. Análisis TF-IDF de las tres muestras*

### *3.1.2 Análisis de bigramas*

En este apartado, utilizamos la técnica de análisis a través de bigramas para representar las diferentes combinaciones de palabras que se presentan en las tres muestras, donde las flechas negras indican una mayor intensidad entre la relación de las palabras.

En la muestra obtenida tras los atentados de Charlie Hebdo, observamos un nodo principal relacionado con el suceso, "charliehebdo". Junto a él, se encuentra el segundo nodo, "jesuischarlie". Ambos relacionados por la palabra "París". En este análisis se muestra cómo las relaciones entre las palabras se centran en el evento como, por ejemplo: "ataque terrorista contra charliehebdo" o "revista satírica semanario" (Figura 16).

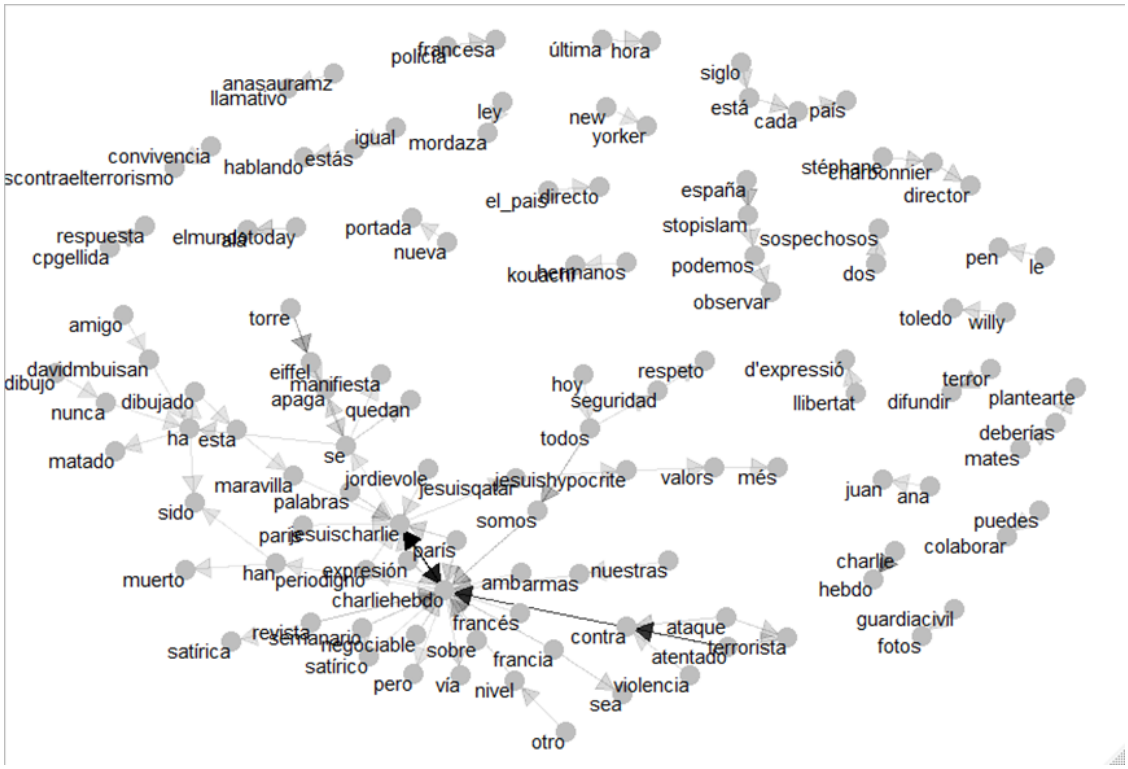


*Figura 16. Análisis de bigramas muestra Charlie Hebdo*

En la muestra obtenida tras los atentados de Bruselas, observamos dos nodos. Por un lado, tenemos el relacionado con el evento y la solidaridad: “prayforbelgium”, “prayfortheworld”, “no al terrorismo”, “jesuisbruxelles”, “atentadobruselas” y por el

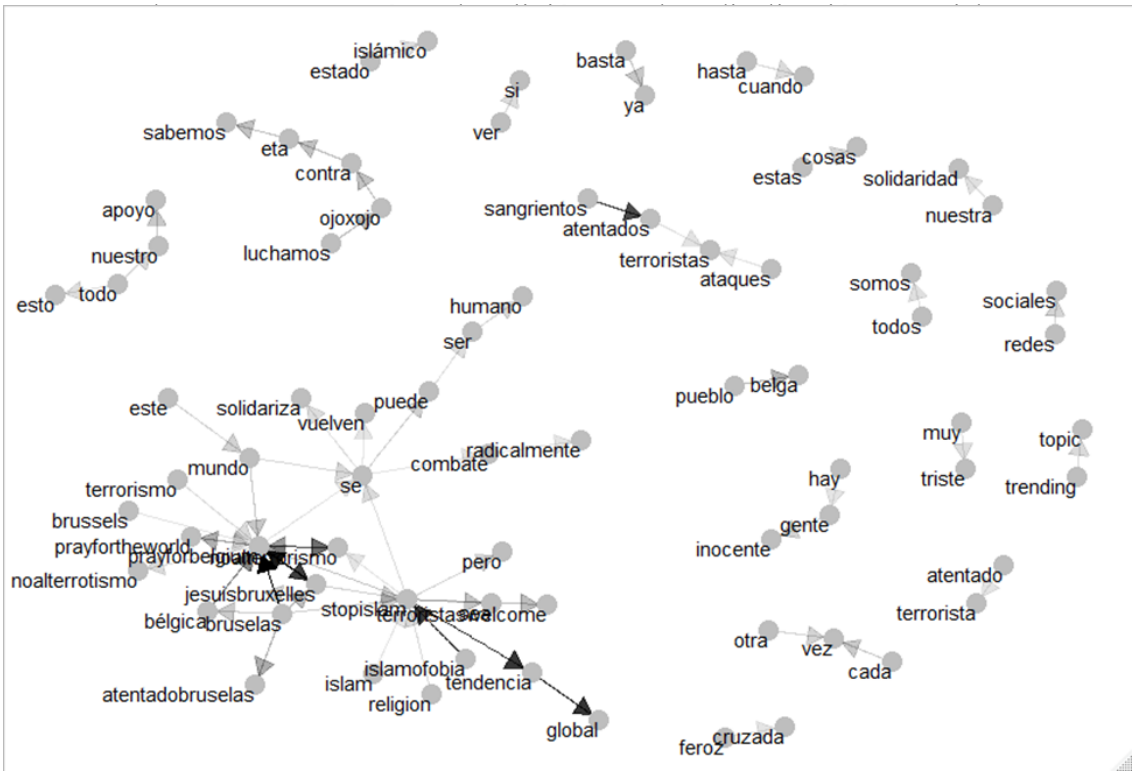


*Figura 17. Análisis de bigramas muestra Bruselas*

En la muestra obtenida tras los atentados de Londres, la relación entre las palabras es mayor, observando dos nodos principales: “londonbridge” y “shelter” ambos relacionados por la palabra “alert”. En el primero, se muestran términos relacionados con el evento: “Hmannella people fleeing boroughmarket after”, “incident”, “incidents”, “terrorist”, “attack”. El segundo nodo se une a un tercero con menor intensidad, en que las palabras se relacionan más con los sentimientos, “drop god heart bless”, “brits to be affected”.

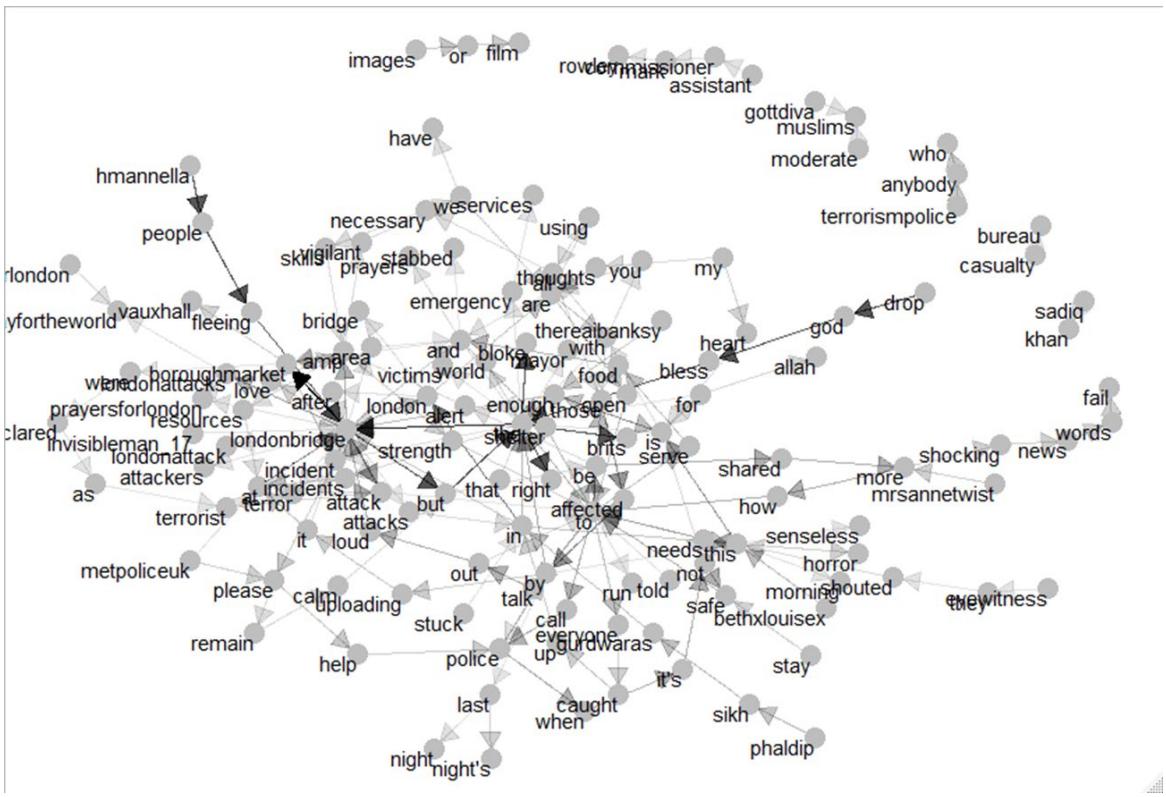


*Figura 18. Análisis de bigramas muestra Londres*

### *3.1.3 Prevalencia de los hashtags del evento, de solidaridad y de radicalización.*

Teniendo en cuenta que cada atentado sucede en un año distinto, es interesante mostrar la prevalencia de los hashtags, en función del evento, para realizar una comparativa entre las tres muestras. En este sentido, en la gráfica de la muestra obtenida tras el atentado de Charlie Hebdo, se observa que, de los tres hashtags, el que más se repite es el del evento, "charliehebdo", con más de 200.000 publicaciones; seguido del de "solidaridad", "jesuischarlie" con casi 75.000 publicaciones y, por último, el de "stopislam" que no llega a 25.000 (Figura 19).

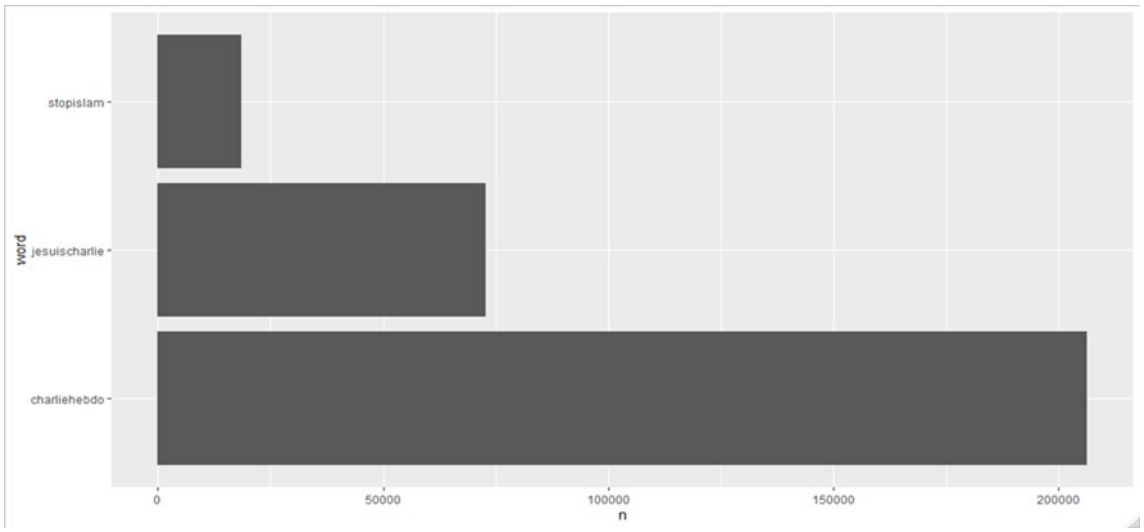


*Figura 19. Cantidad de veces que se repiten los hashtags del evento, de solidaridad y de odio en la muestra de Charlie Hebdo (2015)*

Sin embargo, en la muestra obtenida tras los atentados de Bruselas en 2016, la tendencia da un giro. De los tres hashtags, el que tiene un número mayor de publicaciones es el de odio, "stopislam", con casi 12.000 menciones; seguido del hashtag del evento, "Bruselas", que no llega a 4.000 y, por último, el de solidaridad, "prayforbruselas", con una incidencia mucho menor que el resto (Figura 20).

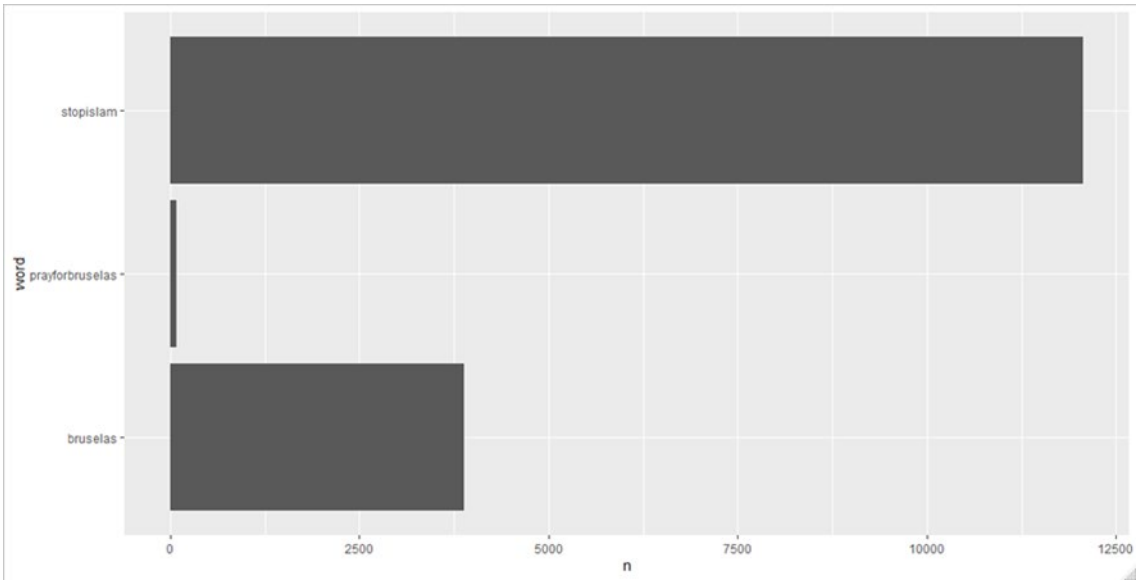


*Figura 20. Cantidad de veces que se repiten los hashtags del evento, de solidaridad y de odio en la muestra de Bruselas (2016)*

Por último, en la muestra obtenida tras los atentados de Londres en 2017, la gráfica vuelve a ser muy parecida a la de Charlie Hebdo. De los tres hashtags, el que más se repite es el del evento, "London", con casi 40.000 menciones; seguido del de solidaridad, "PrayforLondon" con más de 20.000 publicaciones y, por último, el de "stopislam", con una representación mínima (Figura 21).

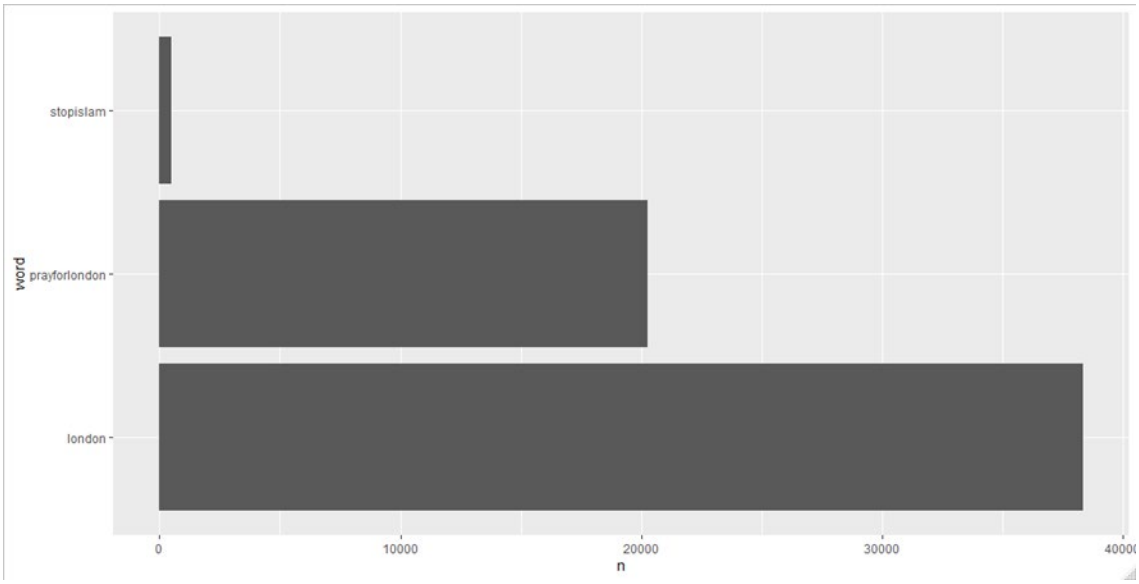


*Figura 21. Cantidad de veces que se repiten los hashtags del evento, de solidaridad y de odio en la muestra de Londres (2017)*

Para finalizar este apartado, se incluyen los resultados derivados de la clasificación realizada entre comunicación neutral y comunicación violenta y discurso de odio de las muestras obtenidas tras los tres atentados terroristas (Tabla 12).

*Tabla 12. Clasificación de los tweets de las tres muestras*

| Evento | Tweets Neutros | % | Tweets CV_DO | % | Total |
|---|---|---|---|---|---|
| Charlie Hebdo | 227.215 | 99,14 | 1.966 | 0,86 | 229.181 |
| Bruselas | 19.390 | 81,28 | 4.473 | 18,72 | 23.863 |
| Londres | 191.392 | 92,26 | 9.488 | 4,74 | 200.880 |

### *3.2. Análisis descriptivo de las variables independientes*

En este apartado se realiza un análisis descriptivo de las variables independientes agrupadas por las categorías propuestas: Anonimato, visibilidad, interacción y estructura.

Anonimato. En la categoría "Anonimato" se analizan las variables que aportan información sobre el titular de la cuenta: *Description* y *Geoenable*.

- Descripción (*Description*)

En la muestra de Charlie Hebdo encontramos que, el 89,25% (*f*=202.781) de los mensajes neutros son emitidos por cuentas en las que se incluye una descripción en la biografía. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio, la descripción aparece en el 93,18% (*f*=1.832) de las cuentas (Tabla 13).

*Tabla 13. Variable Description de la muestra Charlie Hebdo*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 24.434 | 10,75 | 202.781 | 89,25 | 227.215 |
| Cv_Do | 134 | 6,82 | 1.832 | 93,18 | 1.966 |

En la muestra de Bruselas se observa que, el 91,32% (*f*=17.710) de los tweets neutros se ha publicado por cuentas en las que se incluye la descripción en la biografía. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio, la descripción aparece en el 84,52% (*f*=3.777) de las cuentas (Tabla 11).

*Tabla 14. Variable Description de la muestra Bruselas*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 1.684 | 8,68 | 17.710 | 91,32 | 19.394 |
| Cv_Do | 692 | 15,48 | 3.777 | 84,52 | 4.469 |

En la muestra de Londres encontramos que el 82.38% (*f*=157.665) de los tweets neutros se ha publicado por cuentas en las que se incluye la descripción en la biografía. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio, la descripción de la biografía se incluye en el 77,36% (*f*=7.340) de las cuentas (Tabla 15).

*Tabla 15. Variable Description de la muestra Londres*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 33.727 | 17,62 | 157.665 | 82,38 | 191.392 |
| Cv_Do | 2.148 | 22,64 | 7.340 | 77,36 | 9.488 |

Teniendo en cuenta estos datos se puede observar que, exceptuando la muestra obtenida tras el atentado de Charlie Hebdo, la descripción de la cuenta se incluye en mayor porcentaje en las cuentas que han publicado tweets neutros. Así, en la muestra completa de los tres atentados, el 86,34% (*f*=378.156) de los tweets neutros se ha publicado por cuentas en las que se incluye la descripción en la biografía. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio, la descripción de la biografía se incluye en el 81,32% (*f*=12.949) de las cuentas (Tabla 16).

*Tabla 16. Variable Description de la muestra completa*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 59.845 | 13,66 | 378.156 | 86,34 | 438.001 |
| Cv_Do | 2.974 | 18,68 | 12.949 | 81,32 | 15.923 |

- Geolocalización (*Geoenable*)

En la muestra obtenida tras los atentados de Charlie Hebdo encontramos que, de los 227.215 tweet neutros, se incluye la geolocalización del tweet en el 51,09%

(f=116.074). Mientras que, de los 1.966 mensajes clasificados como comunicación violenta y discurso de odio, la geolocalización del tweet se incluye en el 49,95% (f= 982) (Tabla 17).

*Tabla 17. Variable Geoenable de la muestra Charlie Hebdo*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 111.141 | 48,91 | 116.074 | 51,09 | 227.215 |
| Cv_Do | 984 | 50,05 | 982 | 49,95 | 1.966 |

En la muestra obtenida tras los atentados de Bruselas encontramos que, de los 19.394 tweet neutros, se incluye la geolocalización del tweet en el 51,45% (*f*=9.979). Mientras que, de los 4.469 mensajes clasificados como comunicación violenta y discurso de odio, la geolocalización del tweet se incluye en el 42,74% (f= 1.910) (Tabla 18).

*Tabla 18. Variable Geoenable de la muestra Bruselas*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 9.415 | 48,55 | 9.979 | 51,45 | 19.394 |
| Cv_Do | 2.559 | 57,26 | 1910 | 42,74 | 4.469 |

En la muestra obtenida tras los atentados de Londres encontramos que, de los 191.392 tweet neutros, se incluye la geolocalización del tweet en el 43,55% (*f* =83.346). Mientras que, de los 9.488 mensajes clasificados como comunicación violenta y discurso de odio, la geolocalización del tweet se incluye en el 31,96% (*f*=3.032) (Tabla 19).

*Tabla 19. Variable Geoenable de la muestra Londres*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 108.046 | 56,45 | 83.346 | 43,55 | 191.392 |
| Cv_Do | 6.456 | 68,04 | 3.032 | 31,96 | 9.488 |

Teniendo en cuenta estos datos, observamos que la geolocalización de los tweets publicados se presenta en mayor porcentaje, aunque mínimo, en los tweets neutrales. De este modo, de los 438.001 tweet neutrales publicados, el 47,81% (*f*= 209.399) incluyen su geolocalización, mientras que en los tweets de comunicación violenta y discurso es el 37,20% (*f*=5.924) (Tabla 20).

*Tabla 20. Variable Geoenable de la muestra completa*

| **Tweet** | **NO** | | **SÍ** | | **TOTAL** |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 228.602 | 52,20 | 209.399 | 47,81 | 438.001 |
| Cv_Do | 9.999 | 62,80 | 5.924 | 37,20 | 15.923 |

Una vez realizados los análisis descriptivos de las variables pertenecientes a la categoría "Anonimato" se puede observar que los tweets neutrales contienen más información del titular de la cuenta que los tweets de odio. No obstante, es importante destacar que la información aportada en la variable *description* puede no ser real o puede que no aporte información personal sobre el usuario, por lo tanto, aunque se incluya, el anonimato sigue presente.

Visibilidad. En la categoría "Visibilidad" se analizan las variables que hacen visible al usuario de Twitter: *Day_count, Listed_count, Statuses_count, Followers_count, Friends_count y Favourites_count*.

- Días que han pasado entre la publicación del tweet y la creación de la cuenta (*Day_count*)

Respecto a la antigüedad de las cuentas desde las que se emiten los mensajes seleccionados, en la muestra de Charlie Hebdo, se puede observar que las que publican mensajes neutros varían desde los 0 días (es decir, creada el mismo día del evento) hasta 8 años, siendo la media 3 años y medio (M=1.294,26 días; SD=511,95). Sin embargo, la

antigüedad de las cuentas que han publicado mensajes violentos varía entre los 16 días y los 8,7 años, siendo la media de 3,3 años (M=1.227,56 días; SD=505,27) (Tabla 21).

*Tabla 21. Variable Day_count de la muestra Charile Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 1.294,26 | 511,95 | 0 | 3.066 |
| Cv_Do | 1.227,56 | 505,27 | 16 | 2.834 |

En la muestra obtenida tras los atentados de Bruselas se observa que, la antigüedad de las cuentas desde las que se emiten los mensajes neutros varía desde los 4 días hasta los 9,4 años (M=1.268,73 días; SD=756,09). Del mismo modo, la antigüedad de las cuentas que han publicado mensajes clasificados de comunicación violenta o discurso de odio varía entre los 4 días y los 9 años (M=1.260,83; SD=749,68) (Tabla 22).

*Tabla 22. Variable Day_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 1.268,73 | 756,09 | 4 | 3.396 |
| Cv_Do | 1.260,83 | 749,68 | 4 | 3.286 |

Respecto a la antigüedad de las cuentas desde las que se emiten los mensajes seleccionados, en la muestra de Londres, se puede observar que las que publican mensajes neutros varían desde los 0 días (es decir, creada el mismo día del evento) hasta los 10,9 años, siendo la media 4,2 años (M=1.552,81 días; SD=954,15). Así, la antigüedad de las cuentas que han publicado mensajes violentos varía entre los 0 días y los 10,2 años, siendo la media de 3,7 años (M=1.377,21 días; SD=973,65) (Tabla 23).

*Tabla 23. Variable Day_Count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 1.552,81 | 954,15 | 0 | 3.980 |
| Cv_Do | 1.377,21 | 973,65 | 0 | 3.725 |

Tras los análisis descriptivos de esta variable, se destaca que en la muestra de Londres se crearon nuevas cuentas en Twitter el mismo día de los atentados, tanto para publicar mensajes neutros como para los mensajes con contenido violento. Respecto a la antigüedad de las cuentas, no se encuentras diferencias destacadas en las 3 muestras. En general, las cuentas que publican mensajes neutros varían desde los 0 días hasta los 10,9 años, siendo la media de 3,8 años (M=1.406,11; SD=758,83). Las cuentas que emiten mensajes de comunicación violenta y discurso de odio varían entre 0 días y 10,20 años, siendo la media de 3,6 años (M=1.326,07 días; SD=870,65) (Tabla 24).

*Tabla 24. Variable Day_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 1.406,11 | 758,83 | 0 | 3.980 |
| Cv_Do | 1.326,07 | 870,65 | 0 | 3.725 |

- Listas públicas a las que pertenecen los usuarios (*Listed_count*)

En la muestra recogida tras los atentados de Charlie Hebdo, vemos que los usuarios que han publicado mensajes neutros son miembros de 83 listas públicas de media (SD=895,34). Sin embargo, los titulares de los mensajes violentos son miembros de 25,49 listas de media (SD=98,83). Existen usuarios que no pertenecen a ninguna lista y como máximo, los usuarios que han publicado mensajes neutros son miembros de 45.300 listas y los que han publicado mensajes de odio, de 2.782 (Tabla 25).

*Tabla 25. Variable Listed_count de la muestra Charie Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 83,03 | 895,34 | 0 | 45.300 |
| Cv_Do | 25,49 | 98,83 | 0 | 2.782 |

En la muestra recogida tras los atentados de Bruselas, se observa que los usuarios que han publicado mensajes neutros son miembros de 116,66 listas públicas de media

(SD=839,24). Sin embargo, los titulares de los mensajes violentos son miembros de 18 listas de media (SD=75,78) Existen usuarios que no pertenecen a ninguna lista y como máximo, los usuarios que han publicado mensajes neutros son miembros de 51.064 listas y los que han publicado mensajes de odio, de 3.747 (Tabla 26).

*Tabla 26. Variable Listed_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 116,66 | 839,24 | 0 | 51.064 |
| Cv_Do | 18,41 | 75,78 | 0 | 3.747 |

En la muestra recogida tras los atentados de Londres, se muestra que los usuarios que han publicado mensajes neutros son miembros de 59,25 listas públicas de media (SD=485,80). Número muy similar al de los titulares de los mensajes violentos emitidos (M=58,49; SD=379,08). Existen usuarios que no pertenecen a ninguna lista y como máximo, los usuarios que han publicado mensajes neutros son miembros de 90.025 listas y los que han publicado mensajes de odio, de 34.155 (Tabla 27).

*Tabla 27. Variable Listed_count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 59,25 | 485,80 | 0 | 90.025 |
| Cv_Do | 58,49 | 379,08 | 0 | 34.155 |

En el análisis de esta variable se puede observar que los usuarios que han publicado mensajes neutros son miembros de un número mayor de listas públicas que los usuarios que han emitido mensajes con contenido violento. La media general de listas a las que pertenecen los usuarios que han publicado mensajes neutros, es 74,13 (SD=741,88) y

43,17 (SD=297,98) de los usuarios que han emitido mensajes relacionados con la comunicación violenta y el discurso de odio (Tabla 28).

*Tabla 28. Variable Listed_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 74,13 | 741,88 | 0 | 90.025 |
| Cv_Do | 43,17 | 297,98 | 0 | 34.155 |

- Tweets publicados a lo largo de la vida de las cuentas (*Statuses_count*)

Respecto al número de tweets publicados a lo largo de la vida de las cuentas, en la muestra de Charlie Hebdo, los usuarios que han emitido mensajes neutros tienen una media de 18.269,93 (SD=30.629,31). Mientras, los usuarios que han publicado mensajes relacionados con la comunicación violenta y el discurso de odio tienen una media de 19.581, 98 (SD=34.374,05). El número mínimo de mensajes en los usuarios que han publicado tweets neutros es 1. Es decir, que únicamente han publicado un mensaje desde que se unieran a Twitter, siendo el relacionado con el atentado de Bruselas. Sin embargo, el número mínimo de los tweets publicados por los usuarios que han emitido mensajes de odio ha sido 15 (Tabla 29).

*Tabla 29. Variable Statuses_count de la muestra Charile Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 18.269,93 | 30.629,31 | 1 | 416.143 |
| Cv_Do | 19.581,98 | 34.374,05 | 15 | 345.324 |

En la muestra recogida tras los atentados de Bruselas, se observa que el número de tweets emitidos por las cuentas, desde su creación, de los usuarios que han publicado mensajes neutros tiene una media de 27.244,40 (SD=84.810,86). Respecto a los usuarios que han publicado mensajes violentos, la media es de 17,895,79 (SD=35.609,50). El número mínimo del total de los mensajes es 1. Es decir, existen

usuarios de ambas categorías que su primer tweet fue el relacionado con el atentado (Tabla 30).

*Tabla 30. Variable Statuses_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 27.244,40 | 84.810,86 | 1 | 278.7271 |
| Cv_Do | 17.895,79 | 35.609,50 | 1 | 33.4317 |

Respecto al número de tweets publicados a lo largo de la vida de las cuentas, en la muestra de Londres, los usuarios que han emitido mensajes neutros tienen una media de 26.858,94 (SD=65.922,05). Mientras, los usuarios que han publicado mensajes relacionados con la comunicación violenta y el discurso de odio tienen una media de 32.855,06 (SD=58.074,70). El número mínimo del total de los mensajes es 1. Es decir, existen usuarios de ambas categorías que su primer tweet fue el relacionado con el evento (Tabla 31).

*Tabla 31. Variable Statuses_count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 26.858,94 | 65.922,05 | 1 | 7.335.861 |
| Cv_Do | 32.855,06 | 58.074,70 | 1 | 918.840 |

En esta variable cabe destacar que, exceptuando los usuarios analizados en la muestra de Bruselas, los usuarios que han publicado mensajes de odio suelen publicar más tweets que los usuarios que han emitido mensajes neutros. De este modo, la muestra total nos indica que la media de tweets publicados, en general, por usuarios que han publicado mensajes neutros es de 22.420,42 (SD=52.179,01), mientras que los usuarios que han publicado mensajes violentos tienen una media de 27.017,73 (SD=50.613,48) (Tabla 32).

*Tabla 32. Variable Statuses_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 22.420,42 | 52.179,01 | 1 | 7.335.861 |
| Cv_Do | 27.017,73 | 50.613,48 | 1 | 918.840 |

- Seguidores del usuario (*Followers_count*)

En la muestra obtenida tras los atentados de Charlie Hebdo, se observa que los usuarios que han publicado mensajes neutros poseen una media de 5.026, 39 (SD=69.647,14) seguidores. Sin embargo, los usuarios que han publicado mensajes violentos tienen una media de 1.980 (SD=9.168,89). Existen usuarios en ambas categorías que no tienen seguidores (Tabla 33).

*Tabla 33. Variable Followers_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 5.026,39 | 69.647,14 | 0 | 380.1280 |
| Cv_Do | 1.980,67 | 9.168,89 | 0 | 162.866 |

En la muestra obtenida tras los atentados de Bruselas, se observa que los usuarios que han publicado mensajes neutros poseen una media de 19.782,59 (SD=185.199,70) seguidores. Sin embargo, los usuarios que han publicado mensajes violentos tienen una media de 1.506,22 (SD=6.301,40). Existen usuarios en ambas categorías que no tienen seguidores (Tabla 34).

*Tabla 34. Variable Followers_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 19.782,59 | 185.199,70 | 0 | 7.492.682 |
| Cv_Do | 1.506,22 | 6.301,40 | 0 | 164.175 |

En la muestra obtenida tras los atentados de Londres, se observa que los usuarios que han publicado mensajes neutros tienen una media de 3.878,24 (SD=115.687,01) seguidores. Del mismo modo, los usuarios que han publicado mensajes violentos tienen

una media de 3.320,75 (SD=72.099,27). Existen usuarios en ambas categorías que no tienen seguidores (Tabla 35).

*Tabla 35. Variable Followers_count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 3.878,24 | 115.687,01 | 0 | 37.156.858 |
| Cv_Do | 3.320,75 | 72.099,27 | 0 | 6.729.075 |

De los análisis realizados en esta variable se observa que los usuarios que publican mensajes neutros tienen más seguidores que los usuarios que publican mensajes relacionados con la comunicación violenta y el discurso de odio. De este modo, en la muestra completa la media de seguidores de los usuarios que han publicado mensajes neutros es de 5.178,07 (SD=99.464,95), mientras que los usuarios que han emitido mensajes violentos cuentan con una media de 2.646,02 (SD=55.853,18) (Tabla 36).

*Tabla 36. Variable Followers_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 5.178,07 | 99.464,95 | 0 | 37.156,858 |
| Cv_Do | 2.646,02 | 55.853,18 | 0 | 6.729.075 |

- Amigos del usuario (*Friends_count*)

En la muestra obtenida tras los atentados de Charlie Hebdo, se observa que los usuarios que han publicado mensajes neutros poseen una media de 955 (SD=3.365,73) amigos, siendo ligeramente superior para los usuarios que han publicado mensajes violentos, 983,68 (SD=3.396,73). Existen usuarios en ambas categorías que no tienen amigos (Tabla 37).

*Tabla 37. Variable Friends_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Nuetro | 955,40 | 3.365,73 | 0 | 465.665 |
| Cv_Do | 983,68 | 3.396,02 | 0 | 64.470 |

En la muestra obtenida tras los atentados de Bruselas, se observa que los usuarios que han publicado mensajes neutros tienen una media de 1.648,70 (SD=8.835,63) amigos. Mientras que, los usuarios que han publicado mensajes violentos tienen una media de 1.016,18 (SD=4.072,24). Existen usuarios en ambas categorías que no tienen amigos (Tabla 38).

*Tabla 38. Variable Friends_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Nuetro | 1.648,70 | 8.835,63 | 0 | 284.098 |
| Cv_Do | 1.016,18 | 4.072,24 | 0 | 127.693 |

En la muestra obtenida tras los atentados de Londres, se observa que los usuarios que han publicado mensajes neutros tienen una media de 1.489,71 (SD=12.229,41) amigos. Respecto a los usuarios que han publicado mensajes violentos tienen una media de 1.933,93 (5.786,15). Existen usuarios en ambas categorías que no tienen amigos (Tabla 39).

*Tabla 39. Variable Friends_count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 1.489,71 | 12.229,41 | 0 | 3.120.250 |
| Cv_Do | 1.933,93 | 5.786,15 | 0 | 190.542 |

El análisis de esta variable nos indica que, exceptuando la muestra obtenida tras los atentados de Bruselas, el número de amigos de los usuarios que han emitido mensajes con contenido violento es mayor que los que han publicado mensajes neutros. Así, en la muestra completa se observa que la media de amigos es de 1.219,57 (SD=8.646,46) en los usuarios que han publicado mensajes neutros, mientras que en los usuarios que han emitido mensajes de comunicación violenta y discurso de odio es de 1.559,03 (SD=5.121,79) (Tabla 40).

*Tabla 40. Variable Friends_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 1.219,57 | 8.646,46 | 0 | 3.120.250 |
| Cv_Do | 1.559,03 | 5.121,79 | 0 | 190.542 |

- Tweets que el usuario ha marcado como favoritos (Favourites_count)

En la muestra obtenida tras los atentados de Charlie Hebdo, se observa que los usuarios que han publicado mensajes neutros han marcado, de media, 3.343,26 (SD=9.151,27) veces un tweet como favorito durante la existencia de su cuenta. Mientras, los usuarios que han publicado mensajes violentos tienen una media de 4.073,78 (SD=9.539,40). Existen usuarios en ambas categorías que nunca han marcado un tweet como favorito (Tabla 41).

*Tabla 41. Variable Favourites_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 3.343,26 | 9.151,27 | 0 | 272.049 |
| Cv_Do | 4.073,78 | 9.539,40 | 0 | 84.588 |

En la muestra obtenida tras los atentados de Bruselas, se observa que los usuarios que han publicado mensajes neutros han marcado, de media, 4.765,89 (SD=17.802,03) un tweet como favorito durante la existencia de su cuenta. Mientras, los usuarios que han publicado mensajes violentos tienen una media de 5.358,68 (SD=13.723,29). Existen usuarios en ambas categorías que nunca han marcado un tweet como favorito. (Tabla 42).

*Tabla 42. Variable Favourites_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 4.765,89 | 17.802,03 | 0 | 431.750 |
| Cv_Do | 5.358,68 | 13.723,29 | 0 | 196.903 |

En la muestra obtenida tras los atentados de Londres, se observa que los usuarios que han publicado mensajes neutros han marcado, de media, 12.669,20 (SD=28.940,47) veces un tweet como favorito, durante la existencia de su cuenta. Así, los usuarios que han publicado mensajes violentos tienen una media de 18.078,15 (SD=33.967,20). Existen usuarios en ambas categorías que nunca han marcado un tweet como favorito (Tabla 43).

*Tabla 43. Variable Favourites_count de la muestra Londres*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 12.669,20 | 28.940,47 | 0 | 1.171.160 |
| Cv_Do | 18.078,15 | 33.967,20 | 0 | 540.424 |

El análisis de esta variable nos indica que el número de tweets que los usuarios han marcado como favoritos en la vida de su cuenta es mayor en los titulares de cuentas que publican contenido violento. De esta manera, en la muestra completa se observa que la media es de 7.481,38 (SD=21.081,41) en los usuarios que emiten mensajes neutros y 12.779,15 (SD=28.161,85) en los usuarios que publican mensajes relacionados con la comunicación violenta y el discurso de odio (Tabla 44).

*Tabla 44. Variable Favourites_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx. |
|---|---|---|---|---|
| Neutro | 7.481,38 | 21.081,41 | 0 | 1.171.160 |
| Cv_Do | 12.779,15 | 28.161,85 | 0 | 540.424 |

Tras los análisis realizados de las variables que configuran la categoría “visibilidad” observamos que, entre la antigüedad de las cuentas no hay diferencia entre las muestras. Sin embargo, los usuarios que publican mensajes neutros son miembros de un número mayor de listas y tienen más seguidores que los usuarios que publican mensajes relacionados con la comunicación violenta y el discurso de odio, que publican más

tweets, en general, indican en más ocasiones que les gusta un mensaje y tienen más amigos.

Interacción. En la categoría "Interacción" se analizan las variables relacionadas con la interacción entre los usuarios de Twiiter: *Mention_count, Hashtag_count, Link, Retweet_count y Minute_count.*

- Menciones que se han incluido en el mensaje (*Mention_count*)

Respecto al número de menciones que están incluidas los mensajes de la muestra de Charlie Hebdo, se puede observar que en los tweets neutros varían desde las 0 menciones hasta las 10 (a través de la @), siendo la media de 1,04 (SD=0,66). Sin embargo, en los mensajes violentos varía entre las 0 y las 4 menciones, siendo la media de 0,15 (SD=0,47) (Tabla 45).

*Tabla 45. Variable Mention_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 1,04 | 0,66 | 0 | 10 |
| Cv_Do | 0,15 | 0,47 | 0 | 4 |

En la muestra obtenida tras los atentados de Bruselas, el número de menciones que están incluidas en los mensajes se puede observar que en los tweets neutros varían desde las 0 menciones hasta las 10 (a través de la @), siendo la media de 0,15 (SD=0,47). Sin embargo, en los mensajes violentos varía entre las 0 y las 6 menciones, siendo la media de 0,17 (SD=0,49) (Tabla 46).

*Tabla 46. Variable Mention_count de la muestra Bruselas*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 0,15 | 0,47 | 0 | 10 |
| Cv_Do | 0,17 | 0,49 | 0 | 6 |

En la muestra obtenida tras los atentados de Londres se observa que, el número de menciones que están incluidas en los mensajes se puede observar que en los tweets neutros varían desde las 0 hasta las 12 menciones (a través de la @), siendo la media de 0,11 (SD=0,40). Sin embargo, en los mensajes violentos varía entre las 0 y las 6 menciones, siendo la media de 0,11 (SD=0,37) (Tabla 47).

*Tabla 47. Variable Mention_count de la muestra Londres*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 0,11 | 0,40 | 0 | 12 |
| Cv_Do | 0,11 | 0,37 | 0 | 4 |

El análisis de esta variable nos indica que el número de menciones que se han incluido en los tweets publicados es mayor en los titulares de cuentas que publican contenido neutro en la muestra de Charlie Hebdo, menor en la de Bruselas y similar en la de Londres. De esta manera, en la muestra completa se observa que la media es de 0,60 (SD=0,72) en los usuarios que emiten mensajes neutros y 0,13 (SD=0,42) en los usuarios que publican mensajes relacionados con la comunicación violenta y el discurso de odio (Tabla 48).

*Tabla 48. Variable Mention_count de la muestra completa*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 0,60 | 0,72 | 0 | 12 |
| Cv_Do | 0,13 | 0,42 | 0 | 6 |

- Hashtags incluidos en el mensaje (*Hashtag_count*)

En la muestra obtenida tras los atentados de Charlie Hebdo, el número de hashtags que se han incluido en los mensajes neutros varían desde los 0 hasta los 11 (a través de #), siendo la media de 1,35 (SD=0,76). Sin embargo, en los mensajes violentos varía entre 0 y 7 hashtags, siendo la media de 1,43 (SD=0,85) (Tabla 49).

*Tabla 49. Variable Hashtag_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 1,35 | 0,76 | 0 | 11 |
| Cv_Do | 1,43 | 0,85 | 1 | 7 |

En la muestra obtenida tras los atentados de Bruselas, el número de hashtags que se han incluido en los mensajes de ambas categorías varían desde los 0 hasta los 11 (a través de #), siendo la media de 1,68 (SD=1,10) en los mensajes neutros y 1,51 (SD=0,98) en los de comunicación violenta y discurso de odio (Tabla 50).

*Tabla 50. Variable Hashtag_count de la muestra Bruselas*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 1,68 | 1,10 | 0 | 11 |
| Cv_Do | 1,51 | 0,98 | 0 | 11 |

En la muestra obtenida tras los atentados de Londres, el número de hashtags que se han incluido en los mensajes neutros varían desde el 0 hasta los 11 (a través de #), siendo la media de 1,73 (SD=1,09). Mientras, en los mensajes violentos varía entre 0 y 10 hashtags, siendo la media igual que la de los neutros (M=1,73; SD=1,24) (Tabla 51).

*Tabla 51. Variable Hashtag_count de la muestra Londres*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 1,73 | 1,09 | 0 | 11 |
| Cv_Do | 1,73 | 1,24 | 0 | 10 |

El análisis de esta variable nos indica que el número de hashtags incluidos en los tweets publicados es muy similar en las tres muestras. De esta manera, en la muestra completa se observa que la media es de 1,53 (SD=0,95) en los usuarios que emiten mensajes neutros y 1,63 (SD=1,14) en los usuarios que publican mensajes relacionados con la comunicación violenta y el discurso de odio (Tabla 52).

*Tabla 52. Variable Hashtag_count de la muestra completa*

| Tweet | Med | Desv | Min | Max |
|---|---|---|---|---|
| Neutro | 1,53 | 0,95 | 0 | 11 |
| Cv_Do | 1,63 | 1,14 | 0 | 11 |

- Enlaces externos (*Link*)

En la muestra de Charlie Hebdo encontramos que, en el 62,85% (*f*=142.807) de los mensajes neutros se incluye algún tipo de enlace externo. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio, los enlaces externos aparecen en el 19,53% (*f*=384) de los tweets (Tabla 53).

*Tabla 53. Variable Link de la muestra Charlie Hebdo*

| Tweet | NO | | SÍ | | TOTAL |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 84.408 | 37,15 | 142.807 | 62,85 | 227.215 |
| Cv_Do | 1.582 | 80,47 | 384 | 19,53 | 1.966 |

En la muestra obtenida tras los atentados de Bruselas encontramos que, en el 38,21% (*f*=7.410) de los mensajes neutros se incluye algún tipo de enlace externo. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio, los enlaces externos aparecen en el 22,96% (*f*=1.026) de los tweets (Tabla 54).

*Tabla 54. Variable Link de la muestra Bruselas*

| Tweet | NO | | SÍ | | TOTAL |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 11.984 | 61,79 | 7.410 | 38,21 | 19.394 |
| Cv_Do | 3.443 | 77,04 | 1.026 | 22,96 | 4.469 |

En la muestra obtenida tras los atentados de Londres encontramos que, en el 55,97% (*f*=107.125) de los mensajes neutros se incluye algún tipo de enlace externo. Mientras

que, en los mensajes clasificados como comunicación violenta y discurso de odio, los enlaces externos aparecen en el 69,14% (*f*=6.560) de los tweets (Tabla 55).

*Tabla 55. Variable Link de la muestra Londres*

| Tweet | NO | | SÍ | | TOTAL |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 84.267 | 44,03 | 107.125 | 55,97 | 191.392 |
| Cv_Do | 2.928 | 30,86 | 6.560 | 69,14 | 9.488 |

Teniendo en cuenta estos datos podemos observar que, exceptuando la muestra de Londres, los mensajes neutros con contenido externo se encuentran en mayor porcentaje respecto a los de comunicación violenta y discurso de odio. De esta manera, en la muestra general, el 58,75% (*f*=257.342) de los tweets neutros publicados contienen algún enlace externo y el 50% (*f* =7.970) en el caso de los que contienen comunicación violenta o discurso de odio (Tabla 56).

*Tabla 56. Variable Link de la muestra completa*

| Tweet | NO | | SÍ | | TOTAL |
|---|---|---|---|---|---|
| | *f* | % | *f* | % | |
| Neutro | 180.659 | 41,25 | 257.342 | 58,75 | 438.001 |
| Cv_Do | 7.953 | 49,95 | 7.970 | 50,05 | 15.923 |

- Retweets (*Retweet_count*)

Respecto al número de veces que los mensajes analizados se han retweeteado, en la muestra de Charlie Hebdo, se observa que en los tweets neutros la media es de 2.530,38 (SD=5.175), mientras que en los de comunicación violenta y discurso de odio, la media es de 1.220,93 (SD=2.145,36) (Tabla 57).

*Tabla 57. Variable Retweet_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 2.530,38 | 5.175,29 | 0 | 29.810 |
| Cv_Do | 1.220,93 | 2.145,36 | 0 | 5.420 |

Respecto al número de veces que los mensajes analizados se han retweeteado, en la muestra de Bruselas, se observa que en los tweets neutros la media es de 584,14 (SD=4.871,29), mientras que en los de comunicación violenta y discurso de odio, la media es de 195,67 (SD=1.273,91) (Tabla 58).

*Tabla 58. Variable Retweet_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 584,14 | 4.871,29 | 0 | 4.270 |
| Cv_Do | 195,67 | 1.273,91 | 0 | 990 |

Respecto al número de veces que los mensajes analizados se han retweeteado, en la muestra de Londres, se observa que en los tweets neutros la media es de 3.182,57 (SD=5.871,53), mientras que en los de comunicación violenta y discurso de odio, la media es de 1.586,10 (SD=2.260,35) (Tabla 59).

*Tabla 59. Variable Retweet_count muestra Londres*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 3.182,57 | 5.871,53 | 0 | 35.930 |
| Cv_Do | 1.586,10 | 2.260,35 | 0 | 8.276 |

Teniendo en cuenta estos datos podemos observar que los mensajes neutros se retweetean en mayor medida respecto a los de comunicación violenta y discurso de odio. De esta manera, en la muestra general, la media de los tweets neutros que se han retweeteado es de 2.390 (SD=4.198,98) mientras que en el caso de los mensajes violentos la media es de 107,10 (SD=1.910) (Tabla 60).

*Tabla 60. Variable Retweet_count de la muestra completa*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 2.390,81 | 4.189,98 | 0 | 35.930 |
| Cv_Do | 107,10 | 1.910,52 | 0 | 8.276 |

- Minutos transcurridos entre el evento y la publicación del tweet (*Minute_count*)

Respecto a los minutos transcurridos entre el atentado de Charlie Hebdo y la publicación de los mensajes de la muestra se puede observar que los tweets neutros varían desde los 0 minutos (es decir, desde el mismo momento del evento) hasta 5,5 días, siendo la media de 1 día (M=1.515,95 minutos; SD=1.745,59). Del mismo modo, los mensajes referentes a la comunicación violenta y discurso de odio varían entre los 0 minutos hasta los 5,2 días (M=1.341 minutos; SD=1.581,19) (Tabla 61).

*Tabla 61. Variable Minute_count de la muestra Charlie Hebdo*

| **Tweet** | **Med** | **Desv** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| Neutro | 1.515,95 | 1.745,59 | 0 | 7.928 |
| Cv_Do | 1.341,76 | 1.581,19 | 0 | 7.591 |

En la siguiente gráfica muestra la evolución temporal de la repercusión que tuvo en Twitter el ataque terrorista perpetrado a la revista satírica francesa Charlie Hebdo (Figura 22).

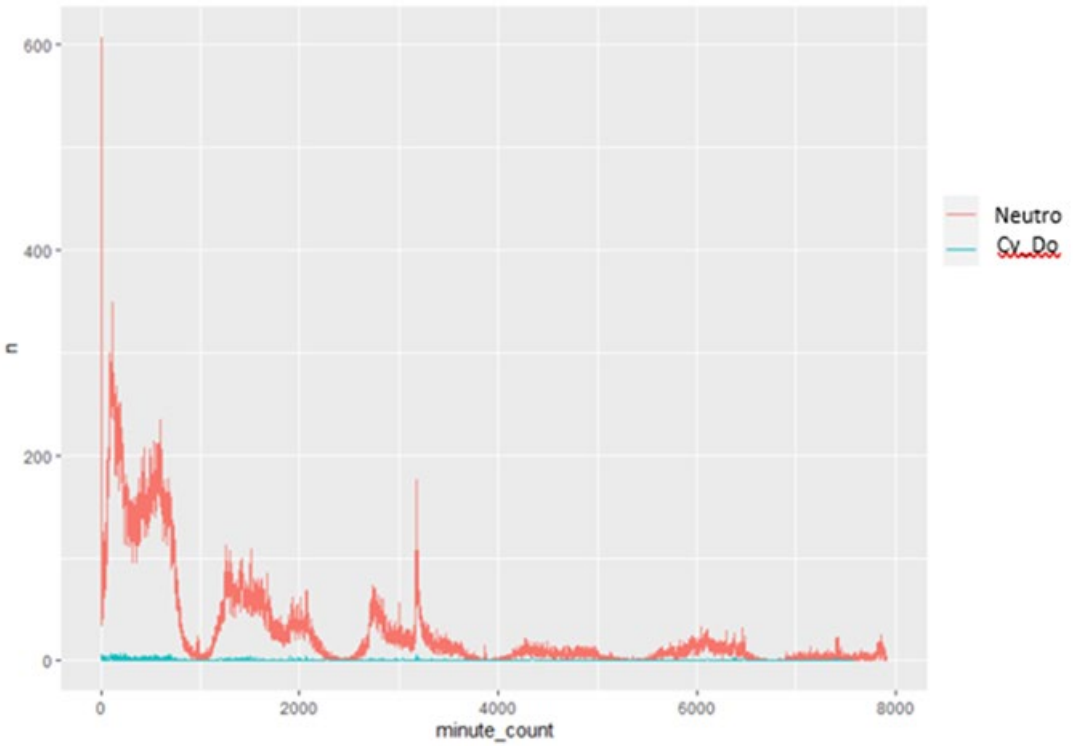


*Figura 22. Análisis temporal muestra Charlie Hebdo*

La línea azul representa la comunicación violenta y el discurso de odio (0,86% de la muestra), mientras que la roja hace referencia a los mensajes neutros. Se observa que los acontecimientos dieron lugar a una gran actividad en la red social, especialmente en los primeros minutos tras el evento. Este hecho responde a un patrón lógico en el que los usuarios reaccionan antes las noticias de los medios de comunicación con comentarios personales, difundiendo los mensajes y creando mayor interacción entre ellos a través de diferentes hashtags como: #CharlieHebdo, #JesuisCharlie y #StopIslam. Conforme va pasando el tiempo existe un descenso en las publicaciones, sin embargo, el jueves 9 de enero se observa un pico, correspondiente con la alarma que se produce al conocer el secuestro de una tiene judía y el atrincheramiento de los hermanos Kouachi en una imprenta. Ese mismo día, el terrorista Coulibaly que asesinó a un policía el día del atentado, mató a cuatro personas y tomó rehenes en un supermercado, antes de ser

abatido. Respecto a la concentración de tweets por días, se observa que el 68,27% de los mensajes se emitió en los primeros días, publicándose el 43,44% de los tweets el mismo día del evento, 7 de enero de 2015, y el 24,83% el día posterior, el miércoles 8 de enero (Tabla 62).

*Tabla 62. Tweets publicados por día muestra Charlie Hebdo*

| Día | Tweets | % |
|---|---|---|
| 7 enero | 99.565 | 43,44 |
| 8 enero | 56.916 | 24,83 |
| 9 enero | 32.468 | 14,17 |
| 10 enero | 11.754 | 5,13 |
| 11 enero | 19.220 | 8,39 |
| 12 enero | 9.258 | 4,04 |

En función de las horas, a través del siguiente reloj aorístico, se aprecia cómo sobre las 11:30 de la mañana del día 7 de enero, aumenta la publicación de mensajes relacionados con el suceso. Debido a la concentración de los tweets, la mayor intensidad se observa en las tres horas siguientes al suceso (Figura 23).

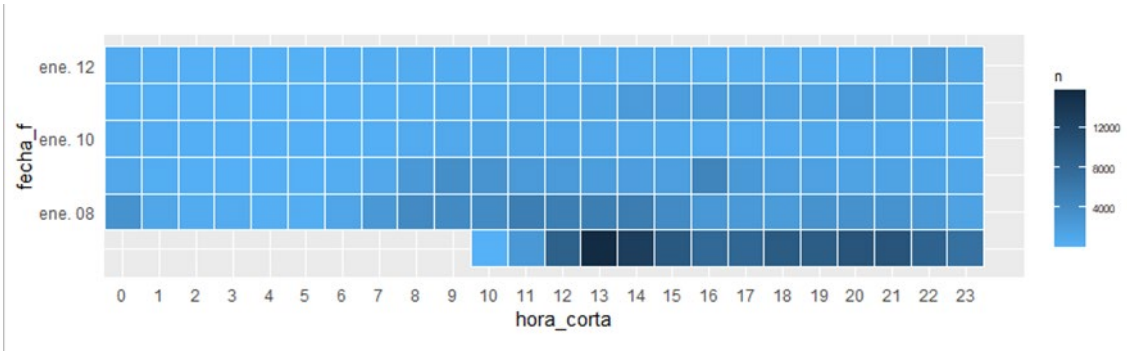


*Figura 23. Mapa de calor muestra Charlie Hebdo*

Respecto a los minutos transcurridos entre el atentado de Bruselas y la publicación de los mensajes de la muestra se puede observar que los tweets neutros varían desde los 10 minutos hasta 6,5 días, siendo la media de 11 horas (M=661,19 minutos;

SD=920,01). Del mismo modo, los mensajes referentes a la comunicación violenta y discurso de odio varían entre los 0 minutos (es decir, en el mismo momento del evento) hasta los 6,4 días (M=882,84 minutos; SD=1.485,15) (Tabla 63).

*Tabla 63.Variable Minute_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín. | Máx |
|---|---|---|---|---|
| Neutro | 661,19 | 920,01 | 10 | 9.448 |
| Cv_Do | 882,84 | 1.485,15 | 0 | 9.315 |

En relación a la evolución temporal de la repercusión que tuvo en Twiiter el atentado de Bruselas, la línea azul representa la comunicación violenta y el discurso de odio (18,72% de la muestra), mientras que la roja hace referencia a los mensajes neutros (Figura 24). Del mismo modo que en Charlie Hebdo, las dos explosiones que tuvieron lugar a las 7:45 de la mañana del 22 de marzo de 2016. en el aeropuerto Zaventem (Bruselas) y la posterior detonación en la estación de metro Maelbeek, a las 9:11, en el centro de la ciudad muy cerca de las principales instituciones europeas generaron, inmediatamente, una gran reacción en los usuarios de la red social.

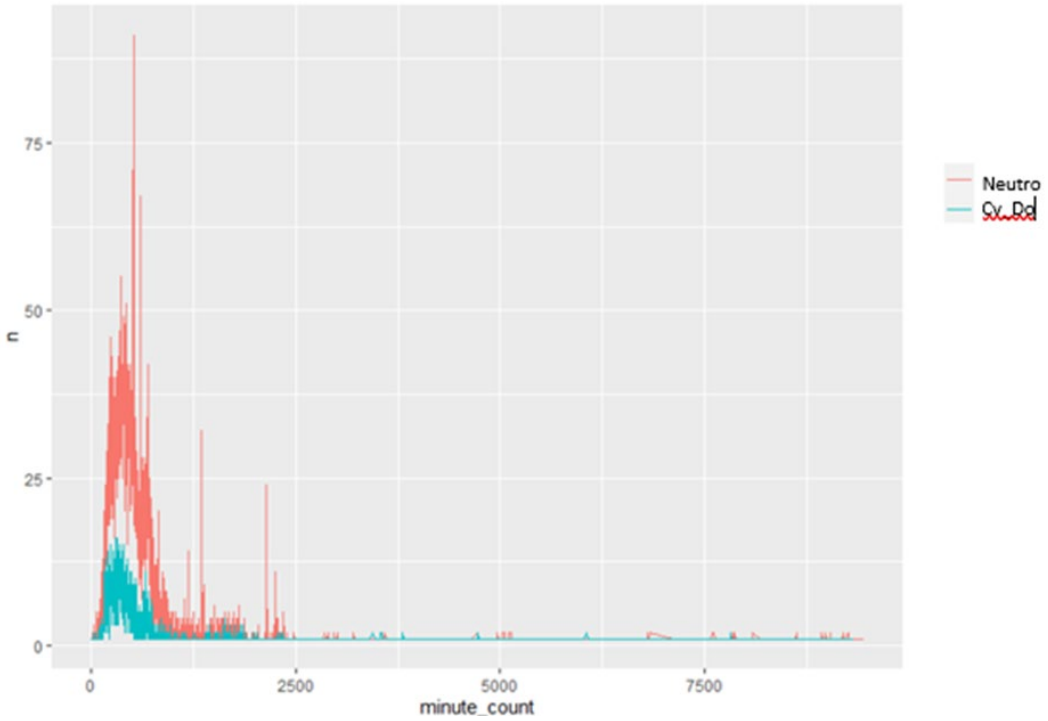


*Figura 24. Análisis temporal muestra Bruselas*

La muestra se recogió una vez sucedió el atentado hasta el 28 de marzo, pero como se puede observar, el 89,22 % de los tweets se publicaron en mismo 22 de marzo de 2015 (Tabla 6)

*Tabla 64. Número de Tweets por día muestra Bruselas*

| Día | Tweets | % |
|---|---|---|
| 22 marzo | 21.290 | 89,22 |
| 23 marzo | 1.812 | 7,59 |
| 24 marzo | 237 | 0,99 |
| 25 marzo | 143 | 0,60 |
| 26 marzo | 112 | 0,47 |
| 27 marzo | 146 | 0,61 |
| 28 marzo | 123 | 0,52 |

En función de las horas, a través del siguiente reloj aorístico, se aprecia cómo sobre las 10:00 de la mañana del 22 de marzo empiezan a aumentar las publicaciones

relacionadas con el suceso, concentrándose el mayor número de ellas, entre las 10:00 y las 16:00 (Figura 25).

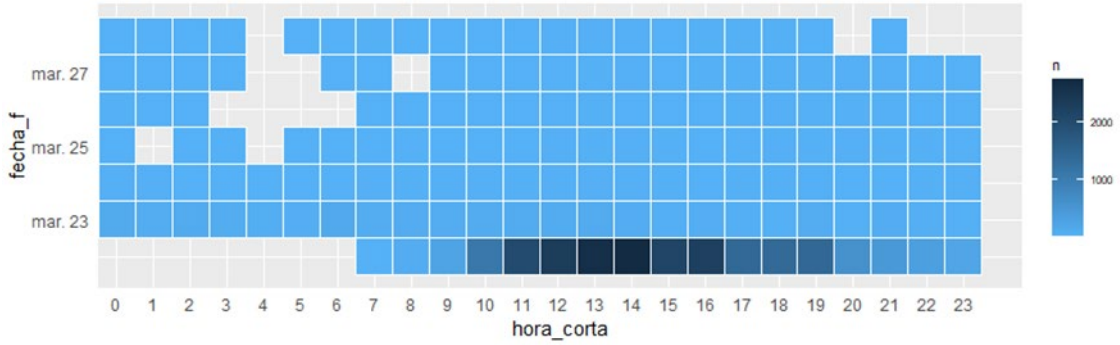


*Figura 25. Mapa de calor muestra Bruselas*

Respecto a los minutos transcurridos entre el atentado de Londres y la publicación de los mensajes de la muestra se puede observar que los tweets neutros varían desde 1 minuto hasta 1,4 días, siendo la media de 11,47 horas (M=688,26 minutos; SD=1362,61). Igualmente, los mensajes referentes a la comunicación violenta y discurso de odio varían desde 1 minuto hasta 1,4 días (M=754,75 minutos; SD=386,32) (Tabla 65).

*Tabla 65. Variable Minute_count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx. |
|---|---|---|---|---|
| Neutro | 688,26 | 362,61 | 1 | 2.134 |
| Cv_Do | 754,75 | 386,32 | 1 | 2.134 |

En relación a la evolución temporal de la repercusión que tuvo en Twiiter estos atentados, la línea azul representa la comunicación violenta y el discurso de odio (4,74% de la muestra), mientras que la roja hace referencia a los mensajes neutros (Figura 26). En este caso la reacción de los usuarios no fue inmediata, sino que se produjo un minuto después de que los terroristas comenzaran a atropellar a los viandantes que paseaban por el puente de Londres e inmediatamente se bajaran del vehículo para apuñalar a las personas que estaban en la zona del mercado, sobre las

22:06 del 3 de junio de 2017. Se llegó al máximo de publicaciones pasada una hora. Después, observamos un descenso que coincide con la noche, volviendo a la actividad por la mañana.

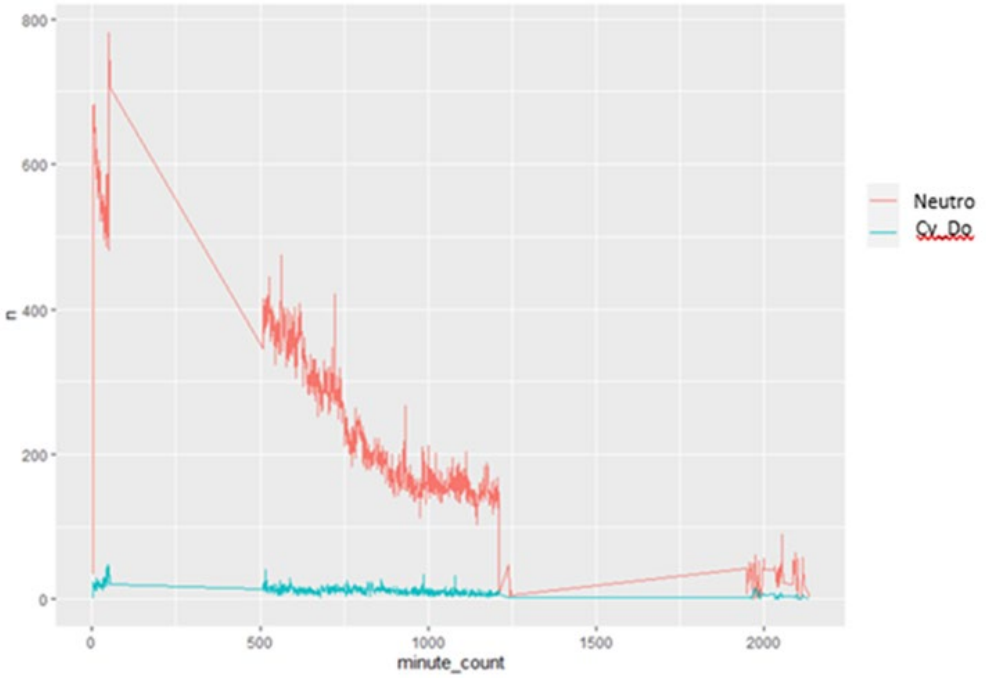


*Figura 26. Análisis temporal muestra Londres*

La muestra se recogió una vez sucedió el atentado hasta el 5 de junio, pero como se puede observar, el 84 % de los tweets se publicaron al día siguiente del evento, el 4 de junio de 2017 (Tabla 66).

*Tabla 66. Tabla de tweet por día muestra Londres*

| Día | Tweets | % |
|---|---|---|
| 3 junio | 30.459 | 15% |
| 4 junio | 167.819 | 84% |
| 5 junio | 2602 | 1% |

En función de las horas, a través del siguiente reloj aorístico, se aprecia de manera más clara, la mayor concentración de las publicaciones entre las 22:30 y las 23:30 de la noche. Del mismo modo, la ausencia de mensajes durante las horas de descanso y la vuelta a la actividad a partir de las 6:30 de la mañana con una concentración mayor sobre las 8 y las 9, volviendo a descender conforme pasan las horas (Figura 27)

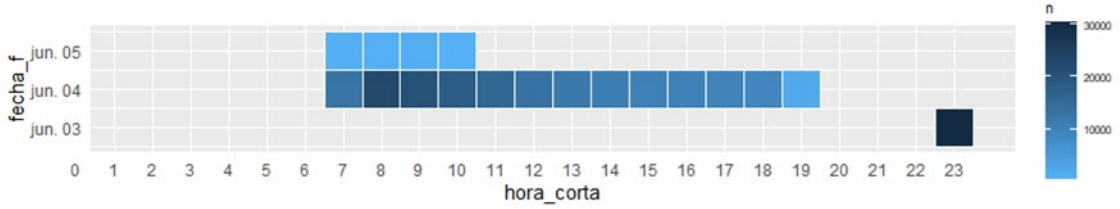


*Figura 27. Mapa de calor muestra Londres*

Una vez realizados los análisis descriptivos de esta variable se destaca que, exceptuando la muestra de Charlie Hebdo, los mensajes neutros se publicaron con una media menor que los de odio. En general, la media del tiempo transcurrido entre el ataque y los mensajes neutros es de 5,7 horas (M=345,33 minutos; SD=450,16) mientras que la de los mensajes violentos es de 6,68 horas (M=701,12 minutos; SD=880,12) (Tabla 67).

*Tabla 67. Variable Minute_count de la muestra completa*

| Tweet | Med | Desv | Mín. | Máx. |
|---|---|---|---|---|
| Neutro | 345,33 | 450,16 | 0 | 9.448 |
| Cv_Do | 701,15 | 880,12 | 0 | 9.315 |

Tras los análisis realizados de las variables que configuran la categoría “Interacción” observamos que, los mensajes neutros incluyen más menciones, se retweetean más, incluyen más enlaces externos y se publican antes que los de comunicación violenta y discurso de odio. Sin embargo, los mensajes de odio emplean más hashtags.

Estructura. En la categoría estructura se analiza el número de caracteres de los mensajes publicados.

- Caracteres del tweet (*Text_count*)

En la muestra obtenida tras los atentados de Charlie Hebdo, se observa que el número de caracteres de los mensajes neutros varía desde 1 hasta 167, siendo la media de 118,89 caracteres (SD=25,43). Sin embargo, en el caso de los mensajes violentos varía entre los 24 y los 144, siendo la media de 107,83 caracteres (SD=29,38) (Tabla 68). Se debe indicar que, aunque el número máximo de caracteres que se podía emplear en Twitter era de 140, la API también recogió los caracteres de enlaces externos, de ahí que el máximo sea superior al número permitido por la red social.

*Tabla 68. Variable Text_count de la muestra Charlie Hebdo*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 111,89 | 25,43 | 1 | 167 |
| Cv_Do | 107,83 | 29,38 | 24 | 144 |

En la muestra obtenida tras los atentados de Bruselas, se observa que el número de caracteres de los mensajes de ambas categorías varía desde 15 hasta 255, siendo la media de 98,38 caracteres (SD=31,78) en el caso de los mensajes neutros y 99,32 (SD=31,28) en el caso de la comunicación violenta y el discurso de odio (Tabla 69).

*Tabla 69. Variable Text_count de la muestra Bruselas*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Nuetro | 98,38 | 31,78 | 15 | 255 |
| Cv_Do | 99,32 | 31,28 | 15 | 255 |

En la muestra obtenida tras los atentados de Londres, se observa que el número de caracteres de los mensajes neutros varía desde los 13 hasta los 255, siendo la media de 102,14 caracteres (SD=30,08). Sin embargo, en el caso de los mensajes violentos varía

entre los 16 y los 255, siendo la media de 108,03 caracteres (SD=30,37) (Tabla 70). Como ya indicó, en el año 2017 Twitter amplió el número de caracteres que se podía emplear en la redacción de los mensajes, de 140 pasaron a 280. El inconveniente para el análisis es que la API no se adaptó a este cambio, teniendo una recogida límite de 255 caracteres.

*Tabla 70. Variable Text_count de la muestra Londres*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Nuetro | 102,14 | 30,08 | 13 | 255 |
| Cv_Do | 108,03 | 30,37 | 16 | 255 |

El análisis de esta variable, teniendo en cuenta las limitaciones descritas, nos indica que el número de caracteres que se emplean para publicar los mensajes de ambas categorías es muy similar. De este modo, en la muestra completa, la media de caracteres en los mensajes neutros es de 107,03 (SD=28,31) mientras que la media de los mensajes violentos es de 105,53 (SD=30,76) (Tabla 71).

*Tabla 71. Variable Text_count de la muestra completa*

| Tweet | Med | Desv | Mín | Máx |
|---|---|---|---|---|
| Neutro | 107,03 | 28,31 | 1 | 255 |
| Cv_Do | 105,56 | 30,76 | 15 | 255 |

### *3.3. Modelo predictivo para diferenciar la comunicación neutra de la comunicación violenta y el discurso de odio*

Para la creación del algoritmo se aplicó un modelo de clasificación basado en la técnica Random Forests (Breiman, 2011), en el que se configuraron diferentes clasificadores para los mensajes. Tras un balanceo al 50% de la muestra de entrenamiento, obtenida tras los atentados de Charlie Hebdo y Bruselas, los clasificadores "aprendieron" del comportamiento de los datos. Ello permitió dividir la

muestra en función de los nodos generados por cada una de las variables incluidas en el modelo. Posteriormente, se validó para el conjunto de variable ambientales de la muestra obtenida tras los atentados de Londres (Tabla 72).

*Tabla 72. Muestras de entrenamiento y validación*

| Clase | Muestra de entrenamiento | Muestra de validación |
|---|---|---|
| Neutral | 6.435 | 191.392 |
| Cv_Do | 6.435 | 9.488 |
| Total | 12.870 | 200.880 |

En función de la información que proporcionan estas variables (i.e., predictores), el algoritmo subdivide la muestra total en muestras homogéneas de menor tamaño para, secuencialmente, clasificar los tweets de la muestra en las dos categorías descritas previamente: contenido neutral y comunicación violenta y discurso de odio.

La precisión del algoritmo para clasificar de manera correcta los mensajes neutros y evitar los falsos positivos es de 0,95. Sin embargo, su precisión en el caso de la comunicación violenta y el discurso de odio es de 0,04. De manera similar, la cantidad de mensajes neutros identificados como tal (i.e., recall) fue de un 0,95 y en el caso de la comunicación violenta y el discurso de odio fue de un 0,05 (Tabla 73).

*Tabla 73. Precisión y Recall del algoritmo*

| Modelo | Precisión | Recall |
|---|---|---|
| Neutral | 0,95 | 0,95 |
| Cv_Do | 0,04 | 0,05 |

En este sentido, observamos que el algoritmo identifica y clasifica mejor el contenido neutral que el contenido relacionado con la comunicación violenta y el discurso de odio. En la siguiente matriz de confusión se muestra información más detallada sobre el número de mensajes clasificados correcta e incorrectamente para el modelo (Tabla 74).

*Tabla 74. Matriz de confusión del modelo*

| Real | Predicción | | Muestra validación |
|---|---|---|---|
| | Neutro | Cv_Do | |
| Neutro | 181.545 | 9.847 | 191.392 |
| Cv_Do | 9.055 | 433 | 9488 |

Además, se realizó la validación de los datos a través de *K-fold cross validation* para evitar que los resultados estuvieran influenciados por la composición del conjunto de datos de entrenamiento. Este proceso se repitió 5 veces; o, dicho de otra manera, se realizaron 5 iteraciones, obteniendo una precisión del 74% mediante el cálculo de la media aritmética de los resultados de cada iteración (Tabla 75).

*Tabla 75. K-fold cross validation*

| *k* | Validación |
|---|---|
| 1 | 0,74 |
| 2 | 0,67 |
| 3 | 0,92 |
| 4 | 0,91 |
| 5 | 0,44 |
| Total | 0,74 |

En cuanto a las variables ambientales de Twitter para clasificar los mensajes, podemos observar la relevancia específica que tiene dentro del modelo (Tabla 76).

*Tabla 76. Importancia de las variables incluidas en el modelo*

| Variable | | Importancia |
|---|---|---|
| Anonimato | | |
| | Description | 0.002082 |
| | Geo_enable | 0.003164 |
| Visibilidad | | |
| | Day_count | 0.026817 |
| | Listed_count | 0.024620 |
| | Statuses_count | 0.022611 |
| | Followers_count | 0.026322 |
| | Friends_count | 0.023225 |
| | Favourites_count | 0.024118 |
| Interacción | | |
| | Mention_count | 0.379393 |
| | Hashtag_count | 0.007610 |
| | Link | 0.100034 |
| | Retweet_count | 0.000079 |
| | Minute_count | 0.313098 |
| Estructura | | |
| | Text_count | 0.046827 |

La puntuación de la importancia se refleja en la proporción de nodos que incluyen una condición impuesta por cada una de las variables enumeradas. En el caso de las variables relacionadas con la interacción del usuario son más importantes para la decisión de salida, mientras que el anonimato tiene un impacto insignificante. Por otro lado, existen tres variables que influyen en el proceso de decisión sobre el resto: La variable *mention_count*, que hace referencia al número de menciones que se registran en el tweet (importancia=0,37); la variable *minute_count*, que mide los minutos que transcurren entre el evento y la publicación del tweet (importancia=0,31) y la variable *link*, que contabiliza la presencia o ausencia de contenido externo en el tweet (importancia=0,1).

Para comprender mejor cuáles son las condiciones específicas que debe cumplir el tweet para ser clasificado por el algoritmo como un mensaje neutral o, por el contrario, una comunicación violenta o discurso de odio, se muestra uno de los árboles de decisión seleccionado de manera aleatoria y transformado en un diagrama de flujo (Figura 28).

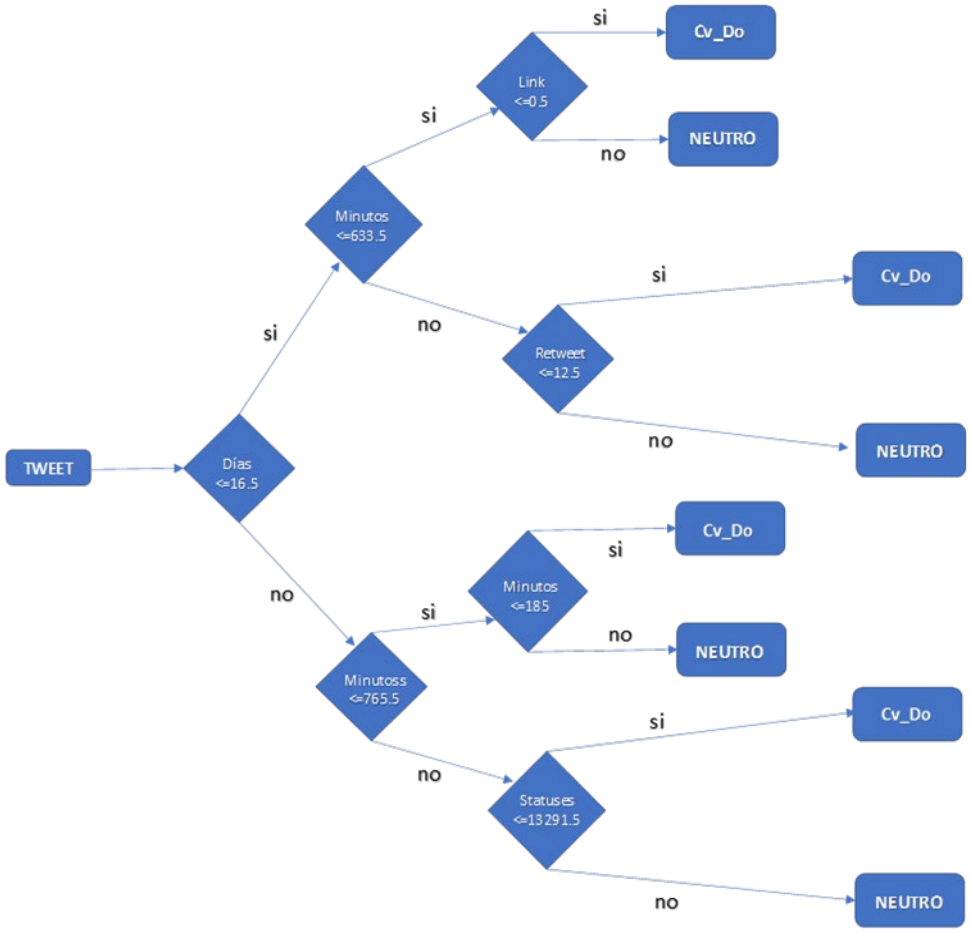


*Figura 28. Diagrama de flujo de uno de los árboles de decisión*

Como se puede observar, los patrones de los metadatos descritos por los mensajes violentos son diferentes de los que se muestran en la comunicación neutral. Este diagrama de flujo muestra algunos contenidos que describen patrones claros y que

pueden ser clasificados usando solo de una a tres variables: Day_count, Minute_count y Retweet_count. Incluso los sellos temporales con baja influencia en el proceso de decisión (Tabla 75), son importantes para definir el contenido de los mensajes.

En resumen, y como se muestra en los datos aportados, es posible definir las condiciones ambientales que deben tener los microlugares de Twitter a fin de diferenciar el tipo de evento que se produce en ellos con certeza. Estas cifras nos permiten interpretar los patrones ambientales que surgen de la combinación de los metadatos asociados a mensajes concretos. Por ejemplo, si un mensaje de la muestra se publicó por una cuenta que tenía de antigüedad menos de 16,5 días, pasadas más de 10,5 horas desde el atentado y ha recibido menos de 12,5 retweets, se clasificó como mensaje de odio: de lo contrario fue clasificado como neutral (Figura 28).

# Capítulo 6

# DISCUSIÓN Y CONCLUSIONES

"The scientific man does not aim at an immediate result. He does not expect that his advanced ideas will be readily taken up. His work is like that of the planter, for the future. His duty is to lay the foundation for those who are to come and point the way"

(Nikola Tesla)

## 1. Discusión

En esta investigación se han utilizado datos de dos muestras de Twitter, en español, para entrenar un modelo predictivo que permita diferenciar la comunicación neutra de la comunicación violenta y el discurso de odio para clasificar automáticamente el contenido de una tercera muestra, en inglés. Para ello, se ha prescindido del análisis semántico y se han utilizado únicamente las variables ambientales del tweet (los metadatos).

Los análisis que se han llevado a cabo han sido, por un lado, descriptivos y por el otro, predictivos. Respecto a los descriptivos, han consistido en el estudio de las tres muestras de mensajes publicados en Twitter, tras los atentados terroristas de Charlie Hebdo (2015), Bruselas (2016) y Londres (2017). Así, como el análisis de las variables

independientes seleccionadas para el estudio. Posteriormente, se ha procedido a la creación de un modelo predictivo para identificar las distintas características ambientales entre la comunicación neutral y la comunicación violenta y el discurso en Twitter.

### *1.1. Características de las muestras obtenidas tras los atentados de Charlie Hebdo, Bruselas y Londres.*

En relación con los análisis semánticos hemos observado que, en la muestra de Charlie Hebdo, las palabras que más se repiten son las relativas al evento y la solidaridad: "libertad", "París", "atentado", "expresión", "Francia" y "hoy". Del mismo modo, en relación con los hashtags seleccionados, el más utilizado fue el del evento, "charliehebdo", seguido del de solidaridad "jesuischarlie" y, por último, el radical, "stopislam". Respecto a las diferentes combinaciones de las palabras obtenidas a través del análisis de bigramas, observamos un nodo principal relacionado con el evento "Charliehebdo" y otro de solidaridad "jesuischarlie", relacionados por la palabra "París". Sobre la base de estos resultados es notorio destacar la mínima expresión de odio mostrado públicamente en Twitter tras este atentado terrorista. Hecho que también queda patente en la clasificación realizada mediante la validación interjueces, donde el 99,14% de los mensajes fueron neutros y únicamente el 0,84% de comunicación violenta y discurso de odio. Estos resultados siguen la misma línea del estudio realizado por Miró-Llinares y Rodríguez-Sala (2016) en el que el 0,8% de la muestra obtenida tras los atentados de Charlie Hebdo (n=282.397) fueron de comunicación violenta y discurso de odio.

Sin embargo, al año siguiente y tras los atentados de Bruselas, la tendencia cambia. El análisis semántico realizado a través del conteo y de la nube de palabras, nos muestra cómo las que más se repiten están relacionadas con el evento, pero también con la religión: “atentadobruselas”, “atentados”, “Bélgica”, “mundo”, “terrorismo”, “religión”, “terroristas”, “musulmanes” e “islamofia”. Quizás, donde se observa mejor es en el análisis de los hashtags seleccionados. El más utilizado fue “stopislam” con casi 12.000 menciones, seguido de “Bruselas” que no alcanza las 4.000 menciones y, por último, “prayforbruselas” con una incidencia mucho menor que el resto. Respecto al análisis de bigramas para comprobar la variedad de las combinaciones entre las palabras, también es manifiesto el contenido radical de mensajes. Sobre todo, en uno de los tres nodos que se centra en la radicalización y en la religión con asociaciones de palabras como “stopislam”, “islamofobia”, “religión” y “tendencia global islamofobia”. Teniendo en cuenta estos resultados, se entiende que el porcentaje de comunicación violenta y discurso de odio en la muestra clasificada mediante la metodología interjueces sea del 18,72% mientras que en los mensajes neutros es del 81,28%.

Por último, el análisis semántico obtenido de la muestra de los atentados sucedidos en Londres en el año 2017 nos indica que el contenido de odio vuelve a descender, ya que las palabras que más se repiten son las relacionas con el evento y el sentimiento de solidaridad: “Boroughmarke”, “people”, “pólice”, “spilling”, “bless”, “brits” y hmannella”. Aquí, es interesante comprobar el análisis de la combinación de palabras a través del análisis de bigramas en la que uno de los nodos principales hace referencia al evento, “londonbridge” relacionando una serie de palabras: “incident”, “incidents”, “terrorist”, “attack” y “Hmannella people fleeing boroughmarket after”. En este sentido, se debe indicar que esta agrupación hace referencia a un mensaje publicado por un usuario de Twitter llamado Howar Mannella, acompañada de una foto en la que un

grupo de jóvenes huían del lugar de los atentados y uno de ellos corría con una cerveza en la mano: "*People fleeing #LondonBridge but the bloke on the right isn't spilling a drop. God Bless the Brits!"*. Tal fue la repercusión de la publicación que, en nuestros análisis de contenido se ha mostrado de manera representativa. Por otro lado, el segundo nodo se une a un tercero con menor intensidad en que las palabras se relacionan más con los sentimientos, "drop god heart bless", "brits to be affected". Sin embargo, donde se observa mucho mejor la tendencia de los tweets es en el análisis de los hashtags seleccionados. El del evento, "London", es el más utilizado con casi 40.000 menciones, seguido del de solidaridad "prayforlondon" con más de 20.000 publicaciones y, por último, el de radicalización "stopislam" con una representación mínima. Así, la información aportada por los análisis descriptivos de la muestra se relaciona con la clasificación obtenida mediante la validación interjueces. Donde el contenido de odio es de un 4,71%, siendo el 95,26% mensajes neutros. En este sentido, observamos que el porcentaje de mensajes de comunicación violenta y discurso de odio es menor que en la muestra de Bruselas, pero mayor que en la de Charlie Hebdo.

Es importante destacar cómo reaccionaron los usuarios de Twitter ante el primer atentado en Europa revindicado por el Estado Islámico. Quizás por esto, la mayoría de los mensajes tras el atentado en la sede del semanario satírico francés, estaban relacionados con el evento y con la solidaridad. Era un suceso nuevo, desconocido para la población. Sin embargo, la tendencia cambia al año siguiente en los atentados de Bruselas, donde se destaca el aumento de los mensajes relacionados con la comunicación violenta y el discurso de odio. Este aumento de mensajes violentos no es constante en el tiempo, puesto que, tras los atentados de Londres, descendió respecto a Bruselas.

Investigaciones posteriores podrían examinar estos cambios en las tendencias de publicación, abordando cuestiones como la repercusión mediática o la cantidad de atentados en un mismo país, entre otros factores que puedan afectar a la respuesta de los usuarios en Twitter.

### *1.2. Características de las variables independientes*

Las variables seleccionadas para la investigación están relacionadas, por un lado, con el anonimato y la visibilidad de las cuentas y por el otro, con la interacción y la estructura de los tweets. De este modo, se obtuvieron 14 variables independientes que se describen a continuación, en función de su relación con la variable dependiente: la comunicación violenta y el discurso de odio.

El análisis de la categoría referente al anonimato, compuesta por las variables *Description* y *Geoenable*, nos muestra la información que el usuario hace pública respecto al apartado biografía de su perfil y la geolocalización de sus tweets. En este sentido, hemos comprobado que los mensajes neutros contenían más información del usuario que los tweets de comunicación violenta y discurso de odio. Ya que, en la muestra total de los tres atentados, el 86,34% de los mensajes neutros fueron publicados por cuentas en las que se incluía la descripción en la biografía. Mientras que, en los mensajes clasificados como comunicación violenta y discurso de odio la descripción se incluía en el 81,32%. En cuanto a la geolocalización de los mensajes analizados de la muestra total, de los 438.001 tweets neutrales, el 47,81% incluían la geolocalización, mientras que en los tweets de comunicación violenta y discurso de odio era el 37,20%.

Una posible explicación para estos hallazgos, en los que no existe mucha diferencia entre unos resultados y otros, sería la falta de información de las variables obtenidas en

las que, únicamente, se indica si existe o no existe descripción en la cuenta y la geolocalización del tweet. Este hecho hace que los resultados no sean del todo adecuados para valorar el anonimato ya que no nos aporta información fiable al respecto. En el caso concreto de que exista una descripción no se asegura que sea real o que realmente aporte información sobre el usuario de la cuenta.

En la categoría relacionada con la visibilidad se analizaron las variables que hacen visible al usuario de Twitter: *Day_count, Listed_count, Statuses_count, Followers_count, Friends_count y Favourites_count*. Respecto a la antigüedad de las cuentas no se encontraron diferencias destacadas en ambas categorías. En general, las cuentas que publicaron mensajes neutros variaban entre los 0 días hasta los 10,9 años, siendo la media de 3,8 años. Y las cuenta que difundían mensajes violentos variaban entre los 0 días y los 10,20 años, siendo la media de 3,6 años. Sin embargo, los usuarios que publicaron mensajes neutros eran miembros de un mayor número de listas (M=74,13) que aquellos que publicaban mensajes clasificados como comunicación violenta y discurso de odio (M=43,17). Además, los usuarios que publicaron mensajes neutros tenían más seguidores (M=5.178,07) que el resto (M=2.646,02). Por su parte, los usuarios que difundieron mensajes violentos y de odio publicaron más tweets, en general (M=27.017,73), que los usuarios que emitieron mensajes neutros (M=22,420,42). También, tenían más amigos (M=1.559,03) que el resto (M=1.219) e indicaron que les gustaba un mensaje en general (M=12.779,15) en más ocasiones que los usuarios que publicaron mensajes neutros (M=7.481,38).

En relación al número de seguidores, Miró-Llinares y Sala-Rodríguez (2016) mostraron cómo los usuarios que tenían entre 100 y 1000 seguidores eran los más representativos tanto en mensajes neutrales como en comunicación violenta y de odio.

Sin embargo, aquellos que tenías entre 1000 y 10.000 seguidores publicaban el 24% de mensajes violentos y de odio y 23,1% neutros.

Estos resultados vienen a apoyar los planteamientos teóricos formulados por Miró-Llinares y Johnson (2018), en los que se incide en la importancia de la visibilidad en este tipo de comportamientos en el ciberespacio, concretamente en Twitter.

En la categoría "Interacción" se analizaron las variables relacionadas con la interacción entre los usuarios de Twiiter: *Mention_count, Hashtag_count, Link, Retweet_count y Minute_count.* Respecto al número de menciones que se incluyeron en los tweets publicados, fue mayor en los titulares de las cuentas que publicaron contenido neutro (M=0,60) respecto de las cuentas que emitieron mensajes violentos y de odio (M=0,13). Los mensajes neutros, incluyeron más enlaces externos (58,75%) que los mensajes violentos y de odio (50%). También, se retweetearon más (M=2.390) que los de odio (M=107,10). Sin embargo, los mensajes relacionados con la comunicación violenta y el discurso de odio incluyeron más hashtags (M=1,63) que los mensajes neutros (M=1,53). Los mensajes con contenido neutro se publicaron antes (M=345,33 minutos) que los mensajes relacionados con la comunicación violenta y el discurso de odio (M=701,12 minutos). Además, los análisis indican que la mayor parte de los tweets se publicaron en el momento de los sucesos. Así, el 43,44% de los mensajes de la muestra de Charlie Hebdo se emitió el mismo día del atento al semanario satírico francés y el 89,22% de los mensajes de la muestra de Bruselas se publicó el 22 de marzo. Por el contrario, el mismo día de los atentados de Londres, se publicó el 15% de los mensajes y el 81% al día siguiente. Estos hallazgos se pueden explicar debido al horario en el que sucedieron los hechos. Los atentados se produjeron por la noche, por lo tanto, tiene sentido que el mayor porcentaje de publicaciones se concentre al día siguiente.

En este sentido y en la línea de otras investigaciones (Miro-Llinares & Rodriguez-Sala, 2016), el mayor flujo de mensajes, tanto neutrales como de violencia y de odio, se concentraron en el momento del ataque, representando el mayor porcentaje en relación total. Estos resultados pueden deberse a que la respuesta emocional a los atentados terroristas varía en función del tiempo transcurrido, que disminuye con el tiempo, dando como resultado más tweets positivos y menos excitados después de que pase tiempo del suceso (Castro-Toledo et al., 2020). En la misma línea, Williams y Burnap (2016) analizaron la reacción de los usuarios de Twitter tras el ataque terrorista de Woolwich en el año 2013. En su investigación mostraron que la difusión de las expresiones de odio online alcanzó su punto máximo poco después del suceso y continuó durante un breve periodo de tiempo. Hecho que coincide con una reacción masiva de los usuarios de la red social inmediatamente después de que el evento ocurra en el espacio físico.

En la categoría estructura se analizó el número de caracteres de los mensajes publicados, siendo muy similar en ambas categorías. En la muestra completa, la media de caracteres en los mensajes neutros fue de 107,03 mientras que la media de los mensajes violentos fue de 105,53.

Investigaciones futuras deberían tener en cuenta los elementos estructurales de los tweets sobre la difusión del contenido y así como la incapacidad de analizarlos de manera correcta debido a la limitación de la API a 255 caracteres.

### *1.3. Características del modelo predictivo para diferenciar la comunicación neutra de la comunicación violenta y el discurso de odio*

Nuestro modelo predictivo se creó utilizando la técnica de clasificación Random Forest, en la que se implementó un algoritmo que estableció diferentes clasificadores

para los tweets. En primer lugar, se realizó un balanceo de datos en la muestra de entrenamiento al 50% a favor de la categoría comunicación violenta y discurso de odio. Esta muestra estaba compuesta por los tweets, en español, obtenidos tras los atentados de Charlie Hebdo y Bruselas. Para la muestra de validación se utilizó la base de datos completa de los tweets obtenidos, en inglés, tras los atentados en Londres. Este hecho implica la capacidad del modelo para poder clasificar los mensajes en la categoría neutral o comunicación violenta y discurso de odio, independientemente del idioma en el que estén escrito, ya que no existe la necesidad de realizar un análisis semántico de cada uno de ellos.

El rendimiento del modelo nos muestra que no todas las variables relacionadas con el anonimato y la visibilidad de los usuarios son criterios adecuados para distinguir si el contenido de un tweet es neutro o, por el contrario, un mensaje relacionado con la comunicación violenta y el discurso de odio.

Concretamente, las variables relacionadas con el anonimato parecen ser irrelevantes a efectos de clasificación, quizás condicionadas por su categorización dicotómica ya que la obtención de información está sesgada hacia variables con un gran número de valores (Quinlan, 1986, citado en Miró-Llinares et al., 2018). En este sentido, pese a que son pocos los estudios empíricos que han profundizado en la influencia del anonimato a nivel individual y pese a que algunos que lo han hecho tangencialmente no extraen resultados concluyentes (Bautista-Ortuño, 2017), existe cierto consenso sobre la existencia entre anonimato en Internet y la conducta delictiva. Así, diferentes investigaciones (Peddinti et al., 2014, 2017b; Xue et al., 2017) han mostrado que los usuarios anónimos en Twitter son más desinhibidos, interactúan más, siguen más cuentas y están más dispuestos a exponer su actividad al público. Sin embargo, identificar el anonimato exclusivamente con los criterios que plantean estas

investigaciones no parece que sea una estrategia adecuada. De este modo, siguiendo la línea de investigación propuesta por Esteve et al. (2019), se deberían implementar nuevas metodologías siguiendo un proceso secuencial y sistemático que graduara de una forma más precisa el anonimato en Twitter.

Además, teniendo en cuenta los resultados y coincidiendo con otras investigaciones (Miró-Llinares et al, 2018), creemos que los metadatos asociados a la visibilidad de la cuenta tampoco son útiles para clasificar los tweets porque estos datos se relacionan con un resultado dicotómico de un tweet concreto, y de esta manera, se puede atribuir incorrectamente características radicales a un microlugar no radical (en este sentido, la cuenta). Es decir, estaríamos equiparando las características de una cuenta que únicamente ha emitido un mensaje violento con otra que difunde sistemáticamente este tipo de mensajes.

Por el contrario, en nuestro modelo, las variables más importantes para clasificar el contenido de un tweet son: *Mention_count, Minute_count* y *link*. Es decir, variables que están estrechamente relacionadas con la interacción generada y la visibilidad del mensaje. Según la teoría, los usuarios que publican mensajes violentos y de odio pretender promover la difusión de sus ideas y pensamientos y, así, incluirían ciertos elementos como los enlaces externos, fomentando el atractivo de los mismos (ElSherief et al., 2018; Suh et al., 2010; Zhang et al., 2018). En este sentido y como argumenta Miró-Llinares (2012), no basta con acceder al ciberespacio para ser visible, ya que la cantidad de usuarios que existe en el mundo hace casi imposible su identificación. Por lo tanto, para conseguir gran difusión del discurso de odio y la comunicación violenta en Twitter, es fundamental que exista una interacción entre los usuarios.

Por otro lado, la estructura del tweet no parece tener mucha relevancia en el modelo para clasificar el contenido del tweet. Aunque, como ya se ha comentado todas las

variables tiene su importancia en el resultado final. Porque, al igual que el diseño arquitectónico de un espacio físico puede condicionar la comisión de delitos en ciertos lugares, en el ciberespacio también. En este sentido, cabe señalar que el paradigma teórico más influyente ha sido la prevención del crimen a través del diseño ambiental o CPTED, el cual enfatiza tres conceptos claves como mediadores entre el espacio físico y la delincuencia: la propiedad; vigilancia y el control de acceso (Cozens et al., 2005). Si trasladamos este modelo al ciberespacio, sería posible emplear las estrategias de CPTED para diseñar microlugares seguros, manipulando determinados elementos para reducir la incidencia de los mensajes violentos y de odio, como por ejemplo, limitando el número de menciones o de hashtags (Moneva, 2020).

Una vez conocida la importancia de las variables en el modelo de clasificación, podemos indicar que los mensajes relacionados con la comunicación violenta y el discurso de odio tienen características ambientas distintas a los mensajes neutros.

Respecto a la precisión del algoritmo para clasificar de manera correcta los mensajes neutros y evitar los falsos positivos es de 95%. Sin embargo, su precisión en el caso de la comunicación violenta y el discurso de odio es del 4%. Además, se realizó la validación de los datos a través de *K-fold cross validation*, obteniendo una precisión del 74%. Sobre la base de los resultados de nuestra investigación, el modelo propuesto clasifica mucho mejor los mensajes neutros que los mensajes identificados como comunicación violenta y discurso de odio. Aun así y teniendo en cuenta la característica particular de nuestra investigación en la que el idioma de la muestra de entrenamiento y validación es diferente, este resultado es muy similar al obtenido en otras investigaciones como la de Burnap y Williams (2015) con un 77% o la de Sharma y sus colaboradores (2018) con un 76%.

## 2. Recapitulación y Conclusiones

El fenómeno de la comunicación violenta y el discurso de odio en Internet plantea una serie de problemas que preocupan a la sociedad, ya que tiene el potencial de causar daño y sufrimiento a los individuos, e incluso puede conducir a desórdenes sociales más allá del ciberespacio (Ross et al., 2017; Waseem & Hovy, 2016). Este fenómeno ha evolucionado del mismo modo que lo han hecho las nuevas formas de interacción y comunicación en el ciberespacio, caracterizándose por su fácil acceso, el tamaño de los potenciales receptores, el anonimato y la instantaneidad (Brown, 2018). Cualquier persona puede tener acceso a la Red (e.g. a través de una wifi pública) permitiéndole aumentar el tamaño de su audiencia en comparación al espacio físico. Además, se pierde el contacto físico que puede permitir al ofensor ver la reacción directa de la víctima, eliminando la posibilidad de minimizar su conducta y aumenta el daño producido a la misma debido a que tiene gran cantidad de público ante la ofensa. Otro de los mayores problemas es la instantaneidad, que genera conductas ofensivas más espontáneas y desconsideradas en Internet (Brown, 2018). Asimismo, este fenómeno fomenta el intercambio de mensajes entre usuarios con la misma ideología, así como el ciberterrorismo y el extremismo (Chetty y Alathur, 2018). Por ello, es necesario desarrollar políticas y métodos para identificar y prevenir estos comportamientos online. También han de proponerse diferentes enfoques que puedan contrarrestar el discurso ofensivo, como la educación, la sensibilización de la sociedad sobre la incitación al odio, el aumento de la tolerancia y la formación más específica de las Fuerzas y Cuerpos de Seguridad en este ámbito. Por otro lado, es importante establecer un marco legal para trazar la línea divisoria entre la libertad de expresión y los contenidos de odio en el ciberespacio y aclarar la responsabilidad de los proveedores de servicios de Internet (Blaya, 2019).

Pero como se ha indicado, el objeto de análisis en esta investigación no es la regulación jurídica y normativa de este fenómeno, sino conocer la prevalencia del fenómeno y analizar las diferencias entre sus características ambientales, de acuerdo con la taxonomía formulada por Miró-Llinares (2016) que aborda el fenómeno desde una perspectiva más amplia.

En este sentido, a lo largo de esta investigación, se han revisado las técnicas automatizas más importantes en la lucha contra la comunicación violenta y el discurso de odio en Internet, las cuales están aumentando en los últimos años. Tanto Warner y Hirschberg (2012) como Burnap y Williams (2015a) fueron pioneros en utilizar clasificadores basados en el aprendizaje automático para detectar este tipo de lenguaje. Por su parte, Djuric et al., (2015) y Mikolov (2013), incorporaron la representación de palabras incrustadas y Nobata et al. (2016) combinó elementos de lenguaje predefinidos e incrustaciones de palabras para entrenar un modelo de regresión. Waseem (2016) usó la regresión logística con N-gramas y características específicas de usuario como género y ubicación. También se han llevado a cabo investigaciones más profundas sobre diferentes clases de lenguaje ofensivo (Davidson et al., 2017). Sin embargo, en esta investigación se ha adoptado un enfoque distinto, basado en la criminología ambiental, en el que el idioma no es óbice para analizar las distintas características de este fenómeno.

Para ello, se ha realizado una revisión de las teorías criminológicas tradicionales en el espacio físico hasta llegar a las teorías de la oportunidad. Así, se han mostrado los primeros postulados realizados por Cohen y Felson (1979) que indicaban los cambios sociales y tecnológicos en las actividades cotidianas de la sociedad y, como consecuencia, la modificación de las oportunidades que tenía un individuo para delinquir. Por su parte, Grabosky (2001) trasladó estos postulados al ciberespacio,

siendo pionero en argumentar que estas oportunidades habían aumentado ya que la posibilidad de convergencia entre el delincuente motivado, el objetivo adecuado y la ausencia del guardián en este nuevo ámbito, era mayor. Posteriormente, diferentes autores (Yar, 2005; Miró-Llinares, 2011) analizaron la aplicación de la teoría de Cohen y Felson al ciberespacio. Así, este nuevo espacio virtual ha modificado la forma de relacionarnos debido a la contracción absoluta de las distancias y su facilidad para comunicarnos (Miró-Llinares, 2011). En este sentido, los ciberlugares son más accesibles y la convergencia física de la que hablaban Cohen y Felson desaparece. Sin embargo, la convergencia debe darse de igual manera, pero ahora puede ser en tiempo real o de manera asincrónica (Reyns et al., 2011). Así, es necesario concretar el concepto de ciberlugar, entendiéndolo como un nodo o área de actividad en el ciberespacio en el que el usuario no se encuentra físicamente ubicado, pero sí puede interactuar (Miró-Llinares & Johnson, 2018). Aunque la investigación empírica sobre la criminología de los ciberlugares está en sus inicios, ya existen trabajos pioneros que muestran su importancia en este ámbito (Moneva, 2020). En este sentido, y como se ha indicado, el análisis de Miró-Llinares y Johnson (2018) es esencial, ya que, gracias a su concepto de ciberlugar, afianza la base teórica para la aplicación de la criminología ambiental del lugar a los delitos llevados a cabo en el ciberespacio.

En esta investigación, las redes sociales se han analizado como ciberlugares donde a los usuarios se les permite emitir cualquier clase de contenido alcanzando a multitud de personas de manera casi inmediata (Silva et al., 2016). De todas las redes sociales, Twitter se caracteriza por ser un entorno de gran comunicación en el que existe una elevada interacción entre sus usuarios. Los contenidos emitidos pueden ser observados y compartidos por un gran número de personas (Marwick & Boyd, 2011), aumentando la visibilidad (Miró-Llinares & Johnson, 2018). Además, algunos usuarios interactúan

de forma anónima, ocultando su identidad a través de nombres falsos o pseudónimos (Peddinti et al., 2014), lo que les permite expresar opiniones de cualquier índole sin temor a ser identificados (Peddinti, et al., 2017) favoreciendo, de esta manera, el anonimato.

Todas estas características hacen que Twitter sea la red social perfecta para difundir mensajes relacionados con la comunicación violenta y el discurso de odio, ya que permanecen fijos en el tiempo, llegan a un público global e incrementan de este modo su lesividad (Miró-Llinares et al., 2018).

En función de estas formulaciones teóricas se ha realizado un estudio empírico para conocer la prevalencia del fenómeno y así analizar las características ambientales de los mensajes difundidos en Twitter. El objetivo final es generar un mayor conocimiento que pueda permitir la identificación y prevención de este fenómeno.

En primer lugar, se obtuvieron los datos primarios de Twitter dentro de un contexto que favoreció la interacción entre los usuarios, así como la difusión de mensajes con una gran carga emocional. Posteriormente se clasificaron, en función de la taxonomía de la comunicación violenta y el discurso de odio (Miró-Llinares, 2016), aplicando una metodología de análisis interjueces para validar la clasificación de los tweets. De este primer análisis se obtuvo que la mayoría de los mensajes eran de carácter neutral, siendo mínima la comunicación violenta y el discurso de odio difundido en Twitter. Se puede observar que en la muestra obtenida tras los atentados de Charlie Hebdo los mensajes de comunicación violenta y discurso de odio no superan el 1%. Del mismo modo, este tipo de mensajes en la muestra obtenida tras los atentados de Londres es de 4,7%. Sin embargo, en Bruselas aumenta, pero sigue siendo un porcentaje pequeño ya que no alcanza el 20%. Estos datos confirman la primera hipótesis planteada ya que la mayoría de la comunicación emitida a través de Twitter es de carácter neutro, por los

que la prevalencia de la comunicación violenta y el discurso de odio es escasa. Es decir, los usuarios de Twitter se centran más en comunicar las noticias, describir los sucesos que se han producido y enviar mensajes solidarios a las víctimas. Aun así, es importante continuar con el análisis del fenómeno, su identificación y prevención porque, como hemos visto, los efectos producidos en las víctimas de mensajes violentos y de odio son preocupantes. De igual forma, los altercados, revueltas o agresiones que se pueden trasladar al espacio físico, fruto de este tipo de comunicación, también se deben tener en cuenta (Burnap y Williams, 2016).

Por otro lado, se han analizado las variables ambientales basadas en los metadatos de los tweets recogidos en las muestras. En este sentido, los usuarios que emitieron mensajes neutros incluyeron más información en la descripción de la biografía del usuario y la geolocalización de los tweets, que los que difundieron mensajes relacionados con la comunicación violenta y el discurso de odio. Respecto al número de listas a las que pertenecen los usuarios, se observa que los tweets neutrales eran emitidos por usuarios que pertenecían a más listas que los usuarios que difundían comunicación violenta y discurso de odio. Al igual que utilizan más menciones, incluyen más enlaces externos y retweetean más que los usuarios que han emitido mensajes relacionados con la comunicación violenta y discurso de odio. Del mismo modo, los usuarios que han emitido mensajes neutros tienen más seguidores que los usuarios que han publicado mensajes violentos y de odio. Por el contrario, los que más publican mensajes en general y tienen más amigos, son los que han emitido comunicación violenta y discurso de odio. Además, le han indicado más veces que les gusta un tweet e incluyen más hashtags que los usuarios que han publicado mensajes neutros.

Por tanto, se observa que las variables ambientales, obtenidas a través de los metadatos, son distintas en la comunicación neutral respecto a las de la comunicación violenta y el discurso de odio, confirmándose así la segunda hipótesis. De este modo se muestra que, la comunicación violenta y el discurso de odio no se distribuye de manera aleatoria, sino que tiene unos patrones determinados que se pueden identificar a través de las variables ambientales, al igual que ocurre con los mensajes neutros.

Por último, se ha elaborado un modelo predictivo a partir de las variables ambientales, obtenidas de los metadatos de los tweets, que nos permite distinguir la comunicación neutra de la comunicación violenta y el discurso de odio. Es decir, nuestro algoritmo, tiene una precisión del 95% para clasificar de manera correcta los mensajes neutros y evitar los falsos positivos. Por lo tanto, también confirmamos la tercera hipótesis mostrando que la criminología ambiental puede aplicarse a los entornos del ciberespacio, introduciendo un nuevo enfoque para sustentar los algoritmos de detección de este fenómeno en Twitter. Así, junto con otras investigaciones similares (Miró-Llinares et al., 2018), se contribuye a proporcionar un nuevo enfoque analítico que abre nuevas líneas de investigación para estudiar las diferentes formas de ciberdelincuencia basadas en los metadatos de los ciberlugares. Es importante destacar que no se debe tener en cuenta todas las variables al construir modelos de predicción, sino que se deberán reducir estos modelos a las variables que estén amparadas por esquemas teóricos fundamentados en la problemática a analizar. De esta manera y como se ha indicado a lo largo de la investigación, este es el proceso que se ha llevado a cabo para seleccionar las variables más adecuadas para construir el modelo predictivo presentado en este estudio.

Además, esta nueva metodología, supera las barreras del lenguaje que van asociadas a un idioma consiguiendo una precisión semejante a la de otros enfoques centrados en el

análisis de los contenidos de los mensajes. Los trabajos que se han desarrollado con un enfoque de análisis semántico y en la sintaxis que emplean búsqueda de palabras clave (Décary-Hétu & Morselli, 2011), bolsas de palabras (Waseem & Hovy, 2016) o análisis de sentimiento (Agarwal & Sureka, 2017) tan solo logran precisiones cercanas al 70% y en nuestro algoritmo se han conseguido un 74%.

Así, con el objetivo de facilitar y reducir las tareas de análisis que realizan los encargados de hacer cumplir la ley y los proveedores de servicio se muestra la eficiencia cuantitativa automatizada por parte del modelo predictivo propuesto. De este modo, se puede reducir la dificultad del análisis de contenido necesario para identificar los mensajes relacionados con la comunicación violenta y el discurso de odio. Además, se completa la eficacia cualitativa al aumentar la capacidad de limitar la atención que prestan las autoridades públicas (e.g la policía) o las entidades privadas a los contenidos que están realmente relacionados con comportamientos dañinos como puede ser, en nuestro caso, la comunicación violenta y el discurso de odio en Twitter. En este sentido, gracias al algoritmo propuesto en nuestra investigación, se pueden excluir de la vigilancia los ciberlugares seguros detectados y aunar los esfuerzos en el resto. Explicado de otro modo y comparándolo con el espacio físico, estamos indicándole a los proveedores de servicios y a las Fuerzas y Cuerpos de Seguridad cuáles son los lugares más seguros y dónde no tienen que patrullar porque en esas zonas no se va a cometer ningún tipo de delitos. Así, pueden dedicar todos los efectivos disponibles a la vigilancia de las zonas donde sí existe riesgo de que se produzca un hecho delictivo.

Por otro lado, debemos indicar que el algoritmo no identifica con gran precisión los mensajes relacionados con la comunicación violenta y el discurso de odio (4%), pero es obvio que cualquier análisis de este tipo nunca va a conseguir la precisión obtenida por la clasificación humana, teniendo en cuenta la naturaleza del lenguaje. Incluyendo

además la falta de acuerdo en la comunidad científica de lo que se considera comunicación violenta y discurso de odio. Además, no podemos olvidar el factor humano en este ámbito. Al final, las conductas llevadas a cabo en el ciberespacio, como la comunicación violenta y el discurso de odio en Twitter, son conductas realizadas por individuos motivados por distintas razones y que, a su vez, afectan a otros individuos o colectivos (Flamand & Décary-Hétu, 2020). Por tanto, la toma de decisiones humanas influye en el curso de la acción. De este modo, aunque se ha demostrado la utilidad de los metadatos en términos de precisión para conseguir la clasificación de la muestra obtenida, se deben tener en cuenta otras circunstancias. Sobre la base de los resultados obtenidos, se propone la fusión de esta metodología con la de enfoques semánticos para conseguir una mejor productividad en la identificación de este fenómeno y así conseguir el objetivo fundamental, su prevención.

## 3. Limitaciones y líneas futura de investigación

En este último apartado del capítulo y como aporte final, se deben indicar una serie de limitaciones identificadas a lo largo del trabajo que permitan avanzar en el conocimiento del fenómeno abordado en investigaciones futuras.

En primer lugar, a pesar de que la muestra total de los tweets analizados es amplia (*N*=453.924), se debe recordar que se divide en tres: la primera obtenida tras los atentados al semanario satírico Charlie Hebdo en 2015 (*N*=229.181), la segunda tras los atentados en el aeropuerto y el metro de Bruselas en 2016 (*N*=23.863) y, la tercera, tras los atentados en el puente y el mercado de Londres en 2017 (*N*=200.880). Lo más adecuado hubiera sido que la muestra de Bruselas se asemejara al resto. En este sentido puedes ser posible que los usuarios que publicaron sus mensajes a raíz de este atentado no utilizaron, del mismo modo como en los otros sucesos, los hashtags seleccionados para obtener la

muestra. Así, en próximas investigaciones se debería tener en cuenta este hecho para conseguir homogeneidad en las muestras, pues al final solo son una representación pequeña del total de mensajes que se lanzaron durante esos días.

Por otro lado, aunque los resultados obtenidos en los análisis descriptivos de las variables independientes, relacionadas con la categoría "anonimato", coinciden con la literatura revisada, en el modelo predictivo parecen ser irrelevantes a efectos de clasificación. En este sentido, únicamente se ha podido valorar la existencia o no de la descripción en la biografía del usuario y la geolocalización del tweet. Por lo tanto, sería interesante establecer mecanismos para determinar si esta es real o no, así como la implementación de metodologías que puedan graduar de una forma más precisa el anonimato en Twitter y estudiar, a su vez, si permiten obtener una mayor precisión en los resultados.

Además, como es sabido, Twitter amplió el número de caracteres que se podían utilizar en la redacción de los mensajes (de 140 a 280). Por lo tanto, el análisis en este aspecto se ha visto limitado ya que, la API únicamente permite la recogida de 255. De este modo, en futuras investigaciones se recomienda tener en cuenta el uso de otras herramientas para analizar los elementos estructurales de los mensajes de manera adecuada debido a esta limitación.

Por último, creemos que sería interesante utilizar el modelo predictivo propuesto para identificar las características diferenciales entre los mensajes neutros y la comunicación violenta y el discurso de odio como base de futuras investigaciones relacionadas con otro tipo de sucesos (e.g relacionados con la política o manifestaciones) en las que se incluyeran más variables ambientales (e.g. utilización de emoticonos o verificación de la cuenta).

# REFERENCIAS BIBLIOGRÁFICAS


Agarwal, S., & Sureka, A. (2017). *Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website*. Retrieved from http://arxiv.org/abs/1701.04931

Aghaei, S., Nematbakhsh, M. A., & Farsani, H. K. (2012). Evolution of the World Wide Web: From Web 1.0 to Web 4.0. *International Journal of Web & Semantic Technology*, *3*(1), 1–10. https://doi.org/10.5121/ijwest.2012.3101

Akers, R. L. (1997). *Criminological theories*. Los Ángeles: Roxbury.

Alcantara, J. (2011). *La neutralidad en La Red, y porqué es una mala idea acabar con ella*. Bilbao-Madrid-Montevideo: Biblioteca de Las Indias.

Ali Rohani, V., & Siew Hock, O. (2009). On Social Network Web Sites: Definition, Features, Architectures and Analysis Tools. *Journal of Computer Engineering*, *1*, 3–11. Retrieved from http://jacr.iausari.ac.ir/article_2392_e8aa31fb1ae5cc322a7686de2fe54d61.pdf

Allen, R. (2008). Factors influencing the usage of Social Networking Websites amongst young , professional South Africans. Univerity of Pretoria.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Antón-Prieto, J. I., & Calderón, P. (2011). Informe sobre las percepciones de seguridad e inseguridad derivadas del uso de las Tecnologías de la Información y la Comunicación (TIC). In *Estudios de la Cátedra de Seguridad Universidad de Salamanca*. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=4029038%0Ahttps://dialnet.unirioja.es/descarga/articulo/4029038.pdf

Armstrong, H. L., & Forde, P. J. (2003). Internet anonymity practices in computer crime. *Information Management & Computer Security*, *11*(5), 209–215. https://doi.org/10.1108/09685220310500117

Assimakopoulos, S., Baider, F. H., & Millar, S. (2017). Online hate speech in the European Union: a discourse-analytic perspective. In *SpringerBriefs in linguistics*. Springer Open.

Awan, I. (2014). Islamophobia and twitter: A typology of online hate against muslims on social media. *Policy and Internet*, *6*(2), 133–150. https://doi.org/10.1002/1944-2866.POI364

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *26th International World Wide Web Conference*, 759–760. https://doi.org/10.1145/3041021.3054223

Baider, F. H. (2017). *Pragmatics lost? Evaluation of hate speech in the EU Code of Conduct*.

Banerjee, T., Yazdavar, A. H., Hampton, A., Purohit, H., Shalin, V. L., & Sheth, A. P. (2012). *Identifying Pragmatic Functions in Social Media Indicative of Gender-Based Violence Beliefs*.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers and Technology*, *24*(3), 233–239. https://doi.org/10.1080/13600869.2010.522323

Bautista Ortuño, R. (2017). ¿Eres un ciberhater? Predictores de la comunicación

violenta y el discurso del odio en Internet. *International E-Journal of Criminal Sciences*, *11*(11), 2.

Beer, D. (2008). Social network(ing) sites...revisiting the story so far: A response to danah boyd & Nicole Ellison. *Journal of Computer-Mediated Communication*, *13*(2), 516–529. https://doi.org/10.1111/j.1083-6101.2008.00408.x

Benevenuto, F., Rodrigues, T., Meeyoung, C., & Almeida, V. (2009). Characterizing User Behavior in Online Social Networks. *IMC'09*.

Benko, A., & Lányi, C. S. (2009). History of artificial intelligence. Second Edition. In *Encyclopedia of Information Science and Technology* (Second, pp. 1759–1762). IGI Global.

Bertillon, A. (1909). *Anthropolige métrique*. Paris: Imprimerie national.

Bishop, J. (2008). Understanding and facilitating the development of social networks in online dating communities: A case study and model. *Social Networking Communities and E-Dating Services: Concepts and Implications*, 266–277. https://doi.org/10.4018/978-1-60566-104-9.ch015

Bishop, J. (2012). The psychology of trolling and lurking: The role of defriending and gamification for increasing participation in online communities using seductive narratives. *Virtual Community Participation and Motivation: Cross-Disciplinary Theories*, 160–176. https://doi.org/10.4018/978-1-4666-0312-7.ch010

Bishop, J. (2013). The effect of de-individuation of the internet troller on criminal procedure implementation: An interview with a hater. *International Journal of Cyber Criminology*, *7*(1), 28–48.

Bishop, J. (2014). Representations of 'trolls' in mass media communication: a review of media-texts and moral panics relating to 'internet trolling. *J. Web Based*

*Communities*, *10*(1), 7–24. https://doi.org/10.1080/14680777.2017.1316755

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior*, *45*(January 2018), 163–172. https://doi.org/10.1016/j.avb.2018.05.006

Blazak, R. (2009). Toward a working definition of hate groups. In *Hate Crimes* (pp. 133–162). London: Greenwood.

Bossler, A. M., & Holt, T. J. (2009). On-Line Activities, Guardianship, and Malware Infection: An Examination of Routine Activities Theory. *International Journal of Cyber Criminology*, *3*(1), 400–420.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Brady, P. Q., Randa, R., & Reyns, B. W. (2016). From WWII to the World Wide Web: A Research Note on Social Changes, Online "Places," and a New Online Activity Ratio for Routine Activity Theory. *Journal of Contemporary Criminal Justice*, *32*(2), 129–147. https://doi.org/10.1177/1043986215621377

Branscomb, A. W. (1995). Anonymity , Autonomy , and Accountability : Challenges to the First Amendment in Cyberspaces. *The Yale Law Journal*, *104*(7), 1639–1679.

Brantingham, P. J. Brantingham, P. L. (1981). Environmental criminology. In *Sage Publications*. Beverly Hills, CA.

Brantingham, P. J. Brantingham, P. L. (1991). Introduction to the 1991 reissue: notes on environmental criminology. *Environmental Criminology*, 1–6.

Brantingham, P., & Brantingham, P. (1995). Criminality of place - Crime generators

and crime attractors. *European Journal on Criminal Policy and Research*, *3*(3), 5–26. https://doi.org/10.1007/BF02242925

Brantingham, P. J., & Brantingham, P. L. (2013). Crime pattern theory. In *Environmental criminology and crime analysis* (pp. 110–116).

Brantingham, P. L., & Brantingham, P. J. (1993). Nodes, Paths and Edges: Considerations on the complexity of crime and the physical environment. *Journal of Environmental Psychology*, *13*(1), 3–28.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA.: Wadsworth International Group.

Breiman, Leo. (2001). Random forests. *Random Forests*, *45*(1), 5–32. https://doi.org/10.1201/9780367816377-11

Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *3*(18), 297–326.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., … Sloan, L. (2013). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, *95*, 96–108. https://doi.org/10.1016/j.techfore.2013.04.013

Burnap, P., & Williams, M. (2014). Hate Speech , Machine Classification and Statistical

Modelling of Information Flows on Twitter : Interpretation and Communication for Policy Decision Making. *Internet, Policy & Politics*, 1–18. https://doi.org/http://dx.doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2015a). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2015b). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0072-6

Canter, D. (1983). The purposive Evaluation of places: a facet approach. *Environment and Behavior*, *15*(6), 659–698.

Capeller, W. (2001). Not such a neat net: Some comments on virtual criminality. *Social & Legal Studies*, *10*, 229–242.

Casado-Riera, C. (2017). *Personalidad y preferencias de uso en las redes sociales en línea* (Univerdidad Ramon Llull). Retrieved from http://www.tdx.cat/handle/10803/409670

Castañeda, L, & Sánchez, M. . (2010). Evolución e historia de las redes sociales. In E. M. S.L. (Ed.), *Aprendizaje con Redes Sociales*. Sevilla.

Castañeda, Linda, González Calatayud, V., & Serrano Sánchez, J. L. (2011). Donde habitan los jóvenes: precisiones sobre un mundo de redes sociales. In F. Martínez & I. Solano (Eds.), *Comunicación y relaciones sociales de los jóvenes en la red*

(pp. 47–63). Alicante: Marfil.

Castañeda, Linda, & Gutiérrez, I. (2010). “Redes sociales y otros tejidos online para conectar personas.” In *Aprendizaje con redes sociales. Tejidos educativos para los nuevos entornos* (pp. 17–39).

Castro-toledo, F. J., Gretenkort, T., Esteve, M., & Miró-Llinares, F. (2020). Fear in 280 characteres: A new approach for evaluation of fear over time in cyberspace. In M. K. N. Vania Ceccato (Ed.), *Crime and Fear in Public Places* (pp. 326–344). https://doi.org/10.1111/j.1753-4887.1966.tb08354.x

Castro, F. J., & Bautista, R. (2019). *Explorando la ofensividad percibida de la comunicación violenta en Internet: un estudio piloto descriptivo*. (740773), 1–12. https://doi.org/10.31235/osf.io/7katg

Cavnar, W. B. (1995). Using An N-Gram-Based Document Representation With A Vector Processing Retrieval Model. *NIST Special Publication*, *3*, 269–278.

Cereceda, J., Francisco, F.-O., Jiménez, S., Herrera, D., Carlos, S., & Ferrés, M. (2018). Informe sobre la evolución de los delitos de odio en España. In *Ministerio del Interior - Secretaría de Estado de Seguridad*.

Chen, Y. (2011). *Detecting offensive language in social medias for protection of adolescent online safety harassment* (The Pennsylvania State University). Retrieved from https://etda.libraries.psu.edu/catalog/12609

Chesney, T., & Su, D. K. S. (2010). The impact of anonymity on weblog credibility. *International Journal of Human Computer Studies*, *68*(10), 710–718. https://doi.org/10.1016/j.ijhcs.2010.06.001

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

Chico, E. (1997). La invarianza en la estructura factorial del raven en grupos de delincuentes y no delincuentes. *Psicothema*, *9*(1), 47–55.

Choi, K. (2008). Computer Crime Victimization and Integrated Theory: An Empirical Assessment. *International Journal of Cyber Criminology*, *2*, 308–333.

Choundhury, N. (2014). World Wide Web and Its Journey from Web 1.0 to Web 4.0. *Journal Of Computer Science and Information Technologies*, *5*(6), 8096–8100. https://doi.org/10.1186/1471-2105-9-82

Citron, Danielle. K. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Citron, Danielle Keats, & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*(4), 1435–1484.

Clarke, R. V. (1999). *Hot Products: Understanding, anticipating and reducing demand for stolen goods*. London: Home Office, Research Devolopment and Statics Directorate.

Clarke, R. V., & Felson, M. (2011). The origins of the routine activity approach and situational crime prevention. In *The origins of American criminology: Advances in criminological theory* (pp. 245–260).

Clarke, Ronald V. (1983). Situational Crime Prevention: Its Theoretical Basis and Practical Scope. *Crime and Justice*, *4*, 225–256. https://doi.org/10.1086/449090

Cleckley, H. (1941). *The Mask of Sanity: An Attempt to Clarify Some Issues About the So-Called Psychopathic Personality*. The C. V. Mosby Company.

Cloward, R. A., & Ohlin, L. E. (1960). *Delinquency and oportunity, a Theory of*

*delinquent gans* (Free Press, Ed.). New York, NY.

Coffey, B., & Woolworth, S. (2004). Destroy the scum, and then neuter their families:" The web forum as a vehicle for community discourse? *Social Science Journal*, *41*(1), 1–14. https://doi.org/10.1016/j.soscij.2003.10.001

Cohen-Almagor, R. (2011). Fighting Hate and Bigotry on the Internet. *Policy & Internet*, *3*(3), 89–114. https://doi.org/10.2202/1944-2866.1059

Cohen-Almagor, R. (2015). Why Confronting the Internet's Dark Side? *Philosophia (United States)*, *45*(3), 919–929. https://doi.org/10.1007/s11406-015-9658-7

Cohen-Almagor, R. (2017). Why Confronting the Internet's Dark Side? *Philosophia*, *45*(3), 919–929. https://doi.org/10.1007/s11406-015-9658-7

Cohen, A. K. (1955). *Delinquent Boys: The culture of the gangs*. Glencoe: Free Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational And Psychological Measurement*, *XX*(1), 37–46.

Cohen, L., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*, 588–608.

Comisión Europea contra el Racismo y la Intolerancia, General Policy Recommendation No. 15, on Combating Hate Speech, CRI (2016) 15, Recital, 3.

Corcoran, A., Marsden, P., Zorbach, T., & Rothlingshofer, B. (2006). Blog marketing. In J. Kirby & P. Marsden (Eds.), *Connected marketing: The viral, buzz and word of mouth revolution* (pp. 148–158). Elsevier.

Cornish, D. B., & Clarke, R. V. (1986). *The Reasoning Criminal: Rational Choice Perspectives on Offending*. New York, NY: Springer-Verlag.

Cornish, D. B., & Clarke, R. V. (2008). The rational choice perspective. In

*Environmental criminology and crime analysis* (pp. 21–47).

Correa, D., Silva, L. ., Mondal, M., Benevenuto, F., & Gummadi, F. . (2015). The many shades of anonymity: Characterizing anonymous social media content. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, 71–80.

Cozens, P. ., Saville, G., & Hillier, D. (2005). Crime Prevention Through Environmental Design (CPTED). *Journal of Property Management*, *23*(5), 328–356. https://doi.org/10.1016/C2012-0-03280-2

Crystal, D. (2006). *Language and the Internet*. Cambridge.

Dadvar, M., Ordelman, R., De Jong, F., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. *Dutch-Belgian Information Retrieval Workshop*, 23–26. Retrieved from http://purl.utwente.nl/publications/79872

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Eleventh International Aaai Conference on Web and Social Media*.

De Alsola, J. V. (2008). Las Redes Sociales satisfacen necesidades básicas. Retrieved from http://www.slideshare.net/Julianalsola/las-redes-sociales-1649666

De Haro, J. . (2010). Servicios de Redes Sociales (I): desenredando la madeja. Retrieved from https://jjdeharo.blogspot.com/2010/07/servicios-de-redes-sociales-i.html

de Marneffe, M.-C., & D. Manning, C. (2008). *Stanford typed dependencies manual*.

Décary-Hétu, D., & Morselli, C. (2011). Gang Presence in Social Network Sites. *International Journal of Cyber Criminology*, *5*(2), 876–890.

Del Pino, C., & Galán, E. (2010). Internet y los nuevos consumidores. El nuevo modelo publicitario. *Cuadernos de Comunicación e Innovación*, *84*, 55–64.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, *1816*, 86–95.

Delgado, R., & Stefancic, J. (2014). Hate Speech in Cyberspace. *Wake Forest Law Review*, *319*.

Di Tullio, B. (1980). *La criminologie : bilan et perspectives.* Paris: A. Pedone.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *Fifth International AAAI Conference on Weblogs and Social Media*.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, *2*(3), 1–30. https://doi.org/10.1145/2362394.2362400

Djuric, N., Wu, H., Radosavljevic, V., Grbovic, M., & Bhamidipati, N. (2015). Hierarchical neural language models for joint representation of streaming documents and their content. *Proceedings of the 24th International Conference on World Wide Web*, 248–255. https://doi.org/10.1145/2736277.2741643

Donath, J. (1995). Idenity and Deception in the Virtual Community. *Communities in Cyberspace*, 27–58. https://doi.org/10.1519/JSC.0b013e3181e4f7a9

Douglas, K. M. (2007). Psychology, discrimination and hate groups online. In *Oxford Handbook of Internet Psychology* (pp. 155–164).

Durkheim, E. (1894). *Les règles de la Méthode Sociologique*. Paris: Les Presses

universitaires de France.

Eck, J. E. (1994). *Drug Markets and Drug Places: A Case-Control Study of the Spatial Structure of Illicit Drug Dealing*. Tesis Doctoral. University of Maryland.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 52–61.

Espinar-Ruiz, E., & González-Río, M. J. (2009). Jóvenes en las redes sociales virtuales. Un análisis exploratorio de las diferencias de género. *Revista Del Centro de Estudios Sobre La Mujer de La Universidad de Alicante*, *14*, 87–105.

Esquivel, Y. (2016). El discurso del odio en la jurisprudencia del Tribunal Europeo de Derechos Humanos. *Revista Mexicana de Derecho Constitucional*, (35).

Esteve, M., Miró, F., & Rabasa, A. (2018). Classification of tweets with a mixed method based on pragmatic content and meta-information. *International Journal of Design and Nature and Ecodynamics*, *13*(1), 60–70. https://doi.org/10.2495/DNE-V13-N1-60-70

Esteve, Z., Moneva, A., & Miró-llinares, F. (2019). Can metadata be used to measure the anonymity of Twitter users ? Results of a Confirmatory Factor Analysis. *International E-Journal of Criminal Sciences*, *13*(2019), 1–16.

Exner, F. (1957). *Biología Criminal en sus rasgos fundamentales*. Barcelona: Bosch.

Eysenck, H. J. (1989). *The Causes and Cures of Criminality*. New York, NY: Plenun Press.

Farias-Hernandez, D., Patti, V., & Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology*, *16*(3), 1–

19. https://doi.org/10.1145/2930663.The

Felson, M. (1995). Those who discourage crime. In *Crime and place* (Vol. 4, pp. 53–66). New York: Criminal Justice Press.

Felson, M. (1998). *Crime and Everyday Life* (2nd ed.). Thousand Oaks, California: Pine Forge Press.

Felson, M. (2006). *Crime and Nature*. Thousand Oaks, California: Sage Publications.

Felson, M., & Boba, R. (2010). *Crime and Everyday Life* (4th ed.). Thousand Oaks, California: SAGE Publications.

Felson, M., & Cohen, L. (1980). Human Ecology and Crime: A Routine Activity Approach. *Human Ecology*, *8*(4), 389–406.

Felson, M., & Cohen, L. (2008). Routine activity theory. In *Environmental criminology and crime analysis* (pp. 70–77).

Felson, M, & Clarke, R. V. (1998). Opportunity makes the thief: practical theory for crime prevention. *Police Research Serie*, *98*, 1–36.

Felson, Marcus. (1986). Linking criminal choices, routine activities, informal control, and criminal outcomes. In *The reasoning criminal* (pp. 119–128). New York: Springer-Verlag.

Fernández Cruz, J. . (2008). La teoría de la elección racional y la crítica a la expansión del Derecho penal: la paradoja de la Política criminal neoliberal. Su aplicación a la delincuencia económica y a los aparatos jerárquicos de la delincuencia organiza. In A. Maíllo Serrano & G. Maisonnave Aller (Eds.), *Intersecciones teóricas en criminología: acción, elección racional y teoría etiológica* (pp. 149–172). Madrid: Dykinson.

Finn, J. (2004). A Survey of Online Harassment at a University Campus. *Journal of Interpersonal Violence*, *19*(4), 468–483. https://doi.org/10.1177/0886260503262083

Flores-Cueto, J. J., Morán-Corzo, J. J., & Rodríguez-Vila, J. J. (2009). Las redes sociales. *Boletín Electrónico de La Unidad de Virtualización Académica*, *1*, 133–133.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3232676

Fowler, J., & Rodd, E. (2017). Web 4.0: The Ultra-Intelligent electronic agent is coming.

Fuchs, C. (2009). *Social networking sites and the surveillance society. A critical case study of the usage of studiVZ, Facebook, and MySpace by students in Salzburg in the context of electronic surveillance.* https://doi.org/ISBN 978-3-200-01428-2

Furnell, S. (2002). *Cybercrime: Vandalizing the information society*. London: Addison-Wesley.

Gagliardone, I. (2014). Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia. *SSRN Electronic Journal*, 43. https://doi.org/10.2139/ssrn.2601792

Gaizauskas, R. (2002). An information extraction perspective on text mining: Tasks, technologies and prototype applications. *Euromap Text Mining Seminar*. Sheffield.

Gambäck, B., & Sikdar, U. K. (2017). *Using Convolutional Neural Networks to Classify Hate-Speech*. (7491), 85–90. https://doi.org/10.18653/v1/w17-3013

García-Pablos de Molina, A. (2008). *Tratado de criminologia* (4th ed.). Valencia: Tirant

lo Blanch.

García Guilabert, N. (2014). Victimización de menores por actos de ciberacoso continuado y actividades cotidianas en el ciberespacio. Retrieved from http://nadir.uc3m.es/alejandro/phd/thesisFinal.pdf%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Universidad+de+murcia#0

Gardner, H. (1995). *Siete Inteligencias. La teoría en la práctica*. Barcelona: Ediciones Páidos Ibérica S.A.

Garofalo, R. (1890). *Criminología: estudio sobre el delito, sobre sus causas y la teoría de la represión*. Madrid: Biblioteca de jurisprudencia, Filosofía e historia. Universidad de Salamanca.

Garrido, V., & Redondo, S. (2013). *Principios de Criminología* (4th ed.). Valencia: Tirant lo Blanch.

Gillin, P. (2007). *The new influencers: A marketer's guide to the new social media*. Linden Publishing.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Glueck, S., & Glueck, E. (1950). Unraveling juvenile delinquency. *Juv. Ct. Judges*, *32*(2).

Glueck, S., & Glueck, E. (1956). *Psyique and delinquency*. New York: New York: Harper.

Goga, O., Perito, D., Lei, H., Teixeira, R., Sommer, R., & Tr-13-002, Łγ. (2013). Large-scale Correlation of Accounts Across Social Networks. *University of*

*California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002*. Retrieved from http://www.icsi.berkeley.edu/pubs/techreports/ICSI_TR-13-002.pdf

Gómez Aguilar, A. (2005). Fronteras electrónicas y nuevas dinámicas transnacionales en Internet. *Revista Internacional de Comunicación Audiovisual, Publicidad y Estudios Culturales*, *1*(3), 39–50.

Goring, C. B. (1913). *The english convict: A statistical Study*. Londres: HSMO.

Gottfredson, M. R., & Hirschi, T. (1990). *A General Theory of Crime*. Standford, California: Standford University Press.

Grabosky, P. (2001). Virtual Criminality: Old Wine in New Bottles? *Social & Legal Studies*, *10*(2), 243–249.

Grabosky, Peter. (2007). The internet, technology, and organized crime. *Asian Journal of Criminology*, *2*, 145–161. https://doi.org/10.1007/s11417-007-9034-z

Graham, G. (1999). *The internet: A philosophical inquiry*. Psychology Press.

Greenawalt, K. (1989). Insults and epithets: Are they protected speech? *Rutgers L. Rev*, *42*(2).

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 468–469.

Griffith, S., & Liyanage, L. (2008). An introduction to the potential of social networking sites in education. *Emerging Technologies Conference*, 18–21. Retrieved from http://ro.uow.edu.au/etc08/9

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4),

5–14. https://doi.org/10.1177/0008125619864925

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361. https://doi.org/10.3102/1076998619832248

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, *6*(2), 215–242. https://doi.org/10.1515/JPLR.2010.011

Hare, R. D. (1970). *Psychopathy: theory and research*. Wiley.

Hartel, P., Junger, M., & Wieringa, R. (2011). Cyber-crime Science= Crime Science + Information Security. In *University of Twente*. https://doi.org/10.1109/ISSA.2014.6950489

Hawdon, J., Oksanen, A., & Räsänen, P. (2014). Victims of online hate groups: American youth's exposure to online hate speech. In M. L. James Hawdon, John Ryan (Ed.), *The Causes and Consequences of Group Violence: From Bullies to Terrorists* (pp. 165–182). Lexington Books.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to Online Hate in Four Nations: A Cross-National Consideration. *Deviant Behavior*, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Henggeler, S. W. (1989). Delinquency in adolescence. In *Sage Publications,*. Thousand Oaks, CA: Sage Publications.

Henson, B., & Reyns, B. W. (2015). The only thing we have to fear is fear itself... and crime: The current state of the fear of crime literature and where it should go next. *Sociology Compass*, *9*(2), 91–103. https://doi.org/10.1111/soc4.12240

Hernández, J., Ramírez, M. ., & Ferri, C. (2004). Introducción a la minería de Datos. In *Pearson. Prentice Hall*.

Herring, S., Job-sluder, K., Scheckler, R., Barab, S., Herring, S., Job-sluder, K., … Herring, S. (2002). Searching for Safety Online : Managing " Trolling " in a Feminist Forum Searching for Safety Online : Managing “ Trolling ” in a Feminist Forum. *The Information Society*, *2243*, 371–384. https://doi.org/10.1080/01972240290108186

Hewitt, S., Tiropanis, T., & Bokhove, C. (2016). The problem of identifying misogynist language on Twitter (and other online social spaces). *Proceedings of the 8th ACM Conference on Web Science*, 333–335.

Hirschi, T. (1969). *Causas de la delincuencia. Berkeley y Los Angeles*. Los Ángeles: University of California Press.

Hollis, M.E., Reynald, D. M., Van Bavel, M., Elffers, H., & Welsh, B. C. (2011). Guardianship for crime prevention: a critical review of the literature. *Crime, Law and Social Change*, *56*(1), 53–70.

Hollis, Meghan E., Felson, M., & Welsh, B. C. (2013). The capable guardian in routine activities theory: A theoretical and conceptual reappraisal. *Crime Prevention and Community Safety*, *15*(1), 65–79. https://doi.org/10.1057/cpcs.2012.14

Holmbom, M. (2015). The YouTuber. A Qualitative Study of Popular Content Creators. *Institutionen För Informatik*, 41. Retrieved from http://www.diva-portal.org/smash/get/diva2:825044/FULLTEXT01.pdf

Holt, T. J., & Bossler, A. M. (2008). Examining the applicability of lifestyle-routine activities theory for cybercrime victimization. *Deviant Behavior*, *30*(1), 1–25. https://doi.org/10.1080/01639620701876577

Hooton, E. A. (1931). *Up from the Ape*. New York: The Macmillan Company.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *Ldv Forum*, *20*(1). https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Howard, E. (2017). Freedom of expression and religious hate speech in Europe. In *Freedom of Expression and Religious Hate Speech in Europe*. https://doi.org/10.4324/9781315277257

Huberman, B. A., Romero, D. M., & Wu, F. (2008). *Social networks that matter Twitter under the microscope*. *14*(1), 1–9. https://doi.org/10.5210/fm.v14i1.2317

Hutchings, A., & Hayes, H. (2009). Routine Activity Theory and Phishing Victimisation: Who Gets Caught in the 'Net'? *Current Issues in Criminal Justice*, *20*(3), 433–452. https://doi.org/10.1080/10345329.2009.12035821

Iab (2019). Estudio anual de redes sociales en España. Retrieved from https://iabspain.es/estudio/estudio-anual-de-redes-sociales-2019/

ILGA. (2016). Hate crime and hate speech. Retrieved from http://www.ilga-europe.org/what-we-do/ouradvocacy-work/hate-crime-hate-speech.

INTECO. (2009). *Estudio sobre la privacidad de los datos personales y la seguridad de la información en las redes sociales online* (Vol. 9). Retrieved from https://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Estudios/est_inteco_redesso_022009.pdf

International, C. (2008). Young People and Social Networking Services. Retrieved from http://www.digizen.org/socialnetworking/downloads/Young_People_and_Social_Net working_Services_full_report.pdf

Jacks, W., & Adler, J. (2015). A Proposed Typology of Online Hate Crime. *Open*

*Access Journal of Forensic Psychology*, *7*, 64–89.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. New York, NY: Random House.

Jaidka, K., Zhou, A., & Lelkes, Y. (2019). Brevity is the soul of Twitter: The constraint affordance and political discussion. *Journal of Communication*, *69*(4), 345–372. https://doi.org/10.1093/joc/jqz023

Jeffery, R. (1977). *Crime prevention through environmental design*. California: Sage Publications.

Jessup, L. M., Connolly, T., & Galegher, J. (1990). The effects of anonymity on GDSS group process with an idea-generating task. *MIS Quarterly: Management Information Systems*, *14*(3), 313–321. https://doi.org/10.2307/248893

Jiménez-Acosta, J. J. (2017). Interpretación del capital erótico en las relaciones sociales mediadas por Tinder. https://doi.org/10.1016/j.sbspro.2015.04.758

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, and Computers*, *31*(3), 433–438. https://doi.org/10.3758/BF03200723

Kaakinen, M., Oksanen, A., & Räsänen, P. (2018). Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior*, *78*, 90–97. https://doi.org/10.1016/j.chb.2017.09.022

Kadiyala, A., & Kumar, A. (2017). Applications of Python to evaluate environmental data science problems. *Environmental Progress and Sustainable Energy*, *36*(6), 1580–1586. https://doi.org/10.1002/ep.12786

Kaiser, G. (1983). *Criminología. Una Introducción a sus fundamentos científicos* (2nd ed.). Madrid: Espasa Calpe.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Katyal, S. K. (2003). The New Surveillance. *Case W. Res. L. Rev*, *54*, 258–271. https://doi.org/10.1007/978-1-137-10929-3_21

Kaufman, G. A. (2015). Odium dicta. Libertad de expresión y protección de grupos discriminados en Internet. In *NASPA Journal* (Vol. 42). https://doi.org/10.1017/CBO9781107415324.004

Keipi, T., Kaakinen, M., Oksanen, A., & Räsänen, P. (2017). Social Tie Strength and Online Victimization: An Analysis of Young People Aged 15–30 Years in Four Nations. *Social Media and Society*, *3*(1), 12. https://doi.org/10.1177/2056305117690013

Keller, E., & Berry, J. (2003). *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy*. Simon and Schuster.

Kent, B. J. (2008). Social Networking Sites : will they survive ? *Nebula*, *5*, 44–51. Retrieved from http://www.nobleworld.biz/images/Kent2.pdf

Koppel, T. (n.d.). Hate Web Sites and the Issue of Free Speech.A, January 13, 1998. *BC News Nightline*.

Kretschmer, E. (1961). *Constitución y carácter: investigaciones acerca del problema de la constitución y de la doctrina de los temperamentos* (3ed ed.). Barcelona: Labor.

Kwak, H., Changhyun, L., Park, H., & Moon, S. (2010). What is Twitter, a Social

Network or a News Media? *International World Wide Web Conference Com-Mittee (IW3C2)*. https://doi.org/10.4321/S0004-05922011000200015

Kwak, H., Chun, H., & Moon, S. (2011). Fragile online relationship: A first look at unfollow dynamics in Twitter. *Proceedings of the SIGCHI Conference on Human Factors in Computing System*, 1091–1100. https://doi.org/10.1145/1978942.1979104

Kwan, G. C. E., & Skoric, M. M. (2013). Facebook bullying: An extension of battles in school. *Computers in Human Behavior*, *29*(1), 16–25. https://doi.org/10.1016/j.chb.2012.07.014

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Labrinidis, A., & Jagadish, H. V. (2017). Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment*, *10*, 2032–2033. https://doi.org/10.14778/3055540

Leets, L., & Giles, H. (1997). Words as weapons - When do they wound?: Investigations of harmful speech. *Human Communication Research*, *24*(2), 260–301. https://doi.org/10.1111/j.1468-2958.1997.tb00415.x

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, *48*(6), 1291–1299. https://doi.org/10.1016/j.jesp.2012.07.002

Leukfeldt, E. R., & Yar, M. (2016). Applying Routine Activity Theory to Cybercrime: A Theoretical and Empirical Analysis. *Deviant Behavior*, *37*(3), 263–280. https://doi.org/10.1080/01639625.2015.1012409

Levin, J., & Mcdevitt, J. (1993). *The rising tide of bigotry and bloodshed: Hate crimes*.

New York: Plenum Press.

Levy, M. P. (2009). *Opportunity, environmental characteristics and crime: An analysis of auto theft patterns*. LFB Scholarly Pub.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Lincoln, S., & Robards, B. (2014). 10 years of Facebook. *New Media and Society*, *16*(7), 1047–1050. https://doi.org/10.1177/1461444814543994

Liu, S., & Forss, T. (2014). Combining N-gram based similarity analysis with sentiment analysis in web content classification. *International Conference on Knowledge Discovery and Information Retrieval*, 530–537. https://doi.org/10.5220/0005170305300537

Liu, S., & Forss, T. (2015). New classification models for detecting hate and violence web content. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *1*, 487–495. https://doi.org/10.5220/0005636704870495

Lockyer, L., & Patterson, J. (2008). Integrating social networking technologies in education: A case study of a formal learning environment. *The 8th IEEE International Conference on Advanced Learning Technologies, ICALT*, 529–533. https://doi.org/10.1109/ICALT.2008.67

Lombroso, C. (1876). *L'uomo delinquente: stuiato in rapporto alla antropología, alla medicina legale ed alle discipline carcerarie*. Milano: Hoepli.

López, C. A., & López, R. M. (2017). Hate Speech, Cyberbullying and Online Anonymity. In S. Assimakopoulos, F. Baider, & S. Millar (Eds.), *Online Hate*

*Speech in the European Union A DiscoursenAnalytic Perspective* (pp. 80–83). Springer Open.

Mahmood, T., & Rohail, K. (2012). Analyzing terrorist incidents to support counter-terrorism - Events and methods. *International Conference on Robotics and Artificial Intelligence, ICRAI 2012*, 149–156. https://doi.org/10.1109/ICRAI.2012.6413382

Mannon, J. M. (1997). Domestic and intimate violence: An application of routine activities theory. *Aggression and Violent Behavior*, *2*(1), 9–24. https://doi.org/10.1016/S1359-1789(96)00023-7

Marañon, G. (1946). *Manual de diagnóstico etiológico*. Madrid: Espasa Calpe.

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. *2018 7th Brazilian Conference on Intelligent Systems*, 61–66. https://doi.org/10.1109/BRACIS.2018.00019

Matsuda, M. J., Laurence III, C. R., Delgado, R., & Williams, K. (1993). *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. New York y London: Routledge.

Matsui, S. (2016). The Challenge to multiculturalism: Hate speech ban in Japan. *UBC Law Review*, *49*(1), 427–484.

McCarthy, J. L. (1982). Metadata Management for Large Statistical Databases. *Eighth International Conference on Very Large Data Bases*, 470–502. México, D.F.

McDevitt, J., & Bennett, S. (2002). Hate crime offenders: An expanded typology. *Journal of Social Issues*, *58*(2), 109–116. https://doi.org/10.4324/9780203446188

McGrath, M. G., & Casey, E. (2002). Forensic psychiatry and the internet: Practical perspectives on sexual predators and obsessional harassers in cyberspace. *Journal of the American Academy of Psychiatry and the Law*, *30*(1), 81–94.

Medina, J. E. (2013). Prevención de la conducción influenciada por medio de los mapas del crimen. Un análisis desde la aplicación de las teorías criminológicas ambientales a la seguridad vial en Elche (Universidad Miguel Hernández).

Medina, J. J. (2011). *Políticas y estrategias de prevención del delito y seguridad ciudadana*. Madrid: Madrid: Edisofer, S. L.

Mehdad, Y., & Tetreault, J. (2016). Do Characters Abuse More Than Words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 299–303. https://doi.org/10.18653/v1/w16-3638

Merodio, J. (2010). *Marketing en redes sociales. Mensajes de empresa para gente selectiva*.

Merton, R. K. (1938). Social Structure and anomie. *American Sociological Review*, *3*(5), 672–682.

Messner, S. F., & Tardiff, K. (1985). The social ecology of urban homicide: An application of the "routine activities" approach. *Criminology*, *23*(2), 241–567.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1389–1399. https://doi.org/10.18653/v1/d16-1146

Miró-Llinares, F. (2012). *El cibercrimen. Fenomenología y criminología de la delincuencia en el ciberespacio*. Madrid: Marcial Pons.

Miró-Llinares, F. (2015). Cibercrimen y vida diaria en el mundo 2.0. In F. Miró-Llinares, J. R. Agustina-Sanllehí, J. . Medina-Sarmiento, & L. Summers (Eds.), *Crimen, oportunidad y vida diaria* (pp. 415–456). Madrid: Dykinson.

Miro-Llinares, F., & Rodriguez-Sala, J. J. (2016). Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design and Nature and Ecodynamics*, *11*(3), 406–415. https://doi.org/10.2495/DNE-V11-N3-406-415

Miró-Llinares, Fernando, Moneva, A., & Esteve, M. (2018). Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. *Crime Science*, *7*(1), 1–12. https://doi.org/10.1186/s40163-018-0089-1

Miró Llinares, F. (2011). La Oportunidad Criminal. *Revista Electrónica de Ciencia Penal y Criminología*, *13*(7), 1–55.

Miró Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet. *IDP. Revista de Internet, Derecho y Política*, (22). https://doi.org/10.7238/idp.v0i22.2975

Miró Llinares, F., & Gómez Bellvís, A. B. (2020). Freedom of Expression in Social Media and Criminalization of Hate Speech in Spain: Evolution, Impact and Empirical Analysis of Normative Compliance and Self-Censorship. *Spanish Journal of Legislative Studies*, (1), 1–42. https://doi.org/10.21134/sjls.v0i1.1837

Miró Llinares, F., & Johnson, S. D. (2018). Cybercrime and Place. *The Oxford Handbook of Environmental Criminology*, *1*, 1–27. https://doi.org/10.1093/oxfordhb/9780190279707.013.39

Mitchell, R. S., Michalski, J. G., & Carbonell, T. M. (2013). *Machine Learning. An*

*artificial intelligence approach*. Berlin: Springer.

Mitchell, W. J. (1995). *City of Bits: Space, Place, and the Infobahn*. https://doi.org/10.2307/1511733

Moneva, A. (2020). *Cyber Places , Crime Patterns , and Cybercrime Prevention : An Environmental Criminology and Crime Analysis approach through Data Science*. Universidad Miguel Hernández.

Moore, D., & Rid, T. (2016). Cryptopolitik and the darknet. *Survival*, *58*(1), 7–38. https://doi.org/10.1080/00396338.2016.1142085

Mossie, Z., & Wang, J.-H. (2018). Social Network Hate Speech Detection for Amharic Language. *Computer Science & Information Technology*, 41–55. https://doi.org/10.5121/csit.2018.80604

Nandhini, B. S., & Sheeba, J. I. (2015). Cyberbullying detection and classification using information retrieval algorithm. *In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology*, 1–5.

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology and People*, *28*(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

Newman, O. (1972). *Defensible Space: Crime Prevention Through Urban Design*. New York: Macmillan.

Ngo, F., & Paternoster, R. (2011). Cybercrime Victimization: An Examination of Individual and Situational Level Factors. *International Journal of Cyber Criminology*, *5*(1), 773.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153. International World Wide Web Conferences.

Nockleby, J. T. (2000). Hate speech. In *Encyclopedia of the American constitution* (pp. 1277–1279).

O'Reilly, T., & Milstein, S. (2011). *The Twitter Book*. O'Reilly.

Oboler, A., & Connelly, K. (2014). Hate speech: A quality of service challenge. *Conference on E-Learning, e-Management and e-Services (IC3e)*, 117–121.

Orihuela, J. L. (2011). *Mundo Twitter: Una guía para comprender y dominar la plataforma que cambió la Red*. Barcelona: Alienta.

Pálmadóttir, J. A., & Kalenikova, I. (2018). *Hate speech: An overview and recommendations for combating it*. Retrieved from http://www.humanrights.is/static/files/Skyrslur/Hatursraeda/hatursraeda-utdrattur.pdf

Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). "On the internet, nobody knows you're a dog": A twitter case study of anonymity in social networks. *Proceedings of the 2014 ACM Conference on Online Social Networks*, 83–93. https://doi.org/10.1145/2660460.2660467

Peddinti, S. T., Ross, K. W., & Cappos, J. (2017). User Anonymity on Twitter. *IEEE Security and Privacy*, *15*(3), 84–87. https://doi.org/10.1109/MSP.2017.74

Perry, B. (2009). *Hate Crimes*. 1200. https://doi.org/10.1177/1043986299015001002

Perry, B., & Olsson, P. (2009). Cyberhate: The globalization of hate. *Information and*

*Communications Technology Law*, *18*(2), 185–199. https://doi.org/10.1080/13600830902814984

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). *Detecting Offensive Language in Tweets Using Deep Learning*. 1–17. https://doi.org/10.1007/s10489-018-1242-y

Poland, B. (2016). *Haters: Harassment, abuse, and violence online*. Universidad of Nebraska Press.

Ponce, I. (2012). Redes sociales-Historia de las redes sociales. *Cuadernos de Información y Comunicación*, *14*, 213–231.

Posner, R. A. (2001). The Speech Market and the Legacy of Schenck. In G. R. S. and L. C. Bollinger (Ed.), *Eternally Vigilant: Free Speech in the Modern Era*. University of Chicago Press.

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2013). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Journal of Applied Business Research*, *29*(1), 195–204. https://doi.org/10.19030/jabr.v29i1.7567

Pratt, T. C., Holtfreter, K., & Reisig, M. D. (2010). Routine Online Activity and Internet Fraud Targeting: Extending the Generality of Routine Activity Theory. *Journal of Research in Crime and Delinquency*, *47*(3), 267–296.

Prieto, E. (2011). *La arquitectura de la ciudad global. Redes, no-lugares, naturaleza*. Madrid: Editorial Biblioteca Nueva.

Prieto Gutiérrez, J. (2011). Herramientas para el análisis y monitoreo en Redes Sociales. *IRIE. International Review of Information Ethics*, *16*(181), 33–40.

Raine, A. (1993). *The Psychopathology of Crime: criminal behavior as a clinical*

*disorder*. San Diego: Gulf Professional Publishing.

Raine, A. (2013). *The Anatomy of Violence: The Biological Roots of Crime*. New York: Pantheon.

Raisi, E., & Huang, B. (2016). *Cyberbullying Identification Using Participant-Vocabulary Consistency*. 46–50. Retrieved from http://arxiv.org/abs/1606.08084

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., & Oksanen, A. (2016). Targets of Online Hate: Examining Determinants of Victimization Among Young Finnish Facebook Users. *Violence and Victims*, *31*(4), 708–725. https://doi.org/10.1891/0886-6708.vv-d-14-00079

Reckless, W. C. (1970). American Criminology. *Criminology*, *8*(1), 4–22.

Rey Martínez, F. (2015). Libertad de expresión y discurso del odio. In *Discurso del odio y racismo líquido* (p. 189). Alcalá de Henares (Madrid): Universidad de Alcalá de Henares.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, *47*(1), 239–268. https://doi.org/10.1007/s10579-012-9196-x

Reyns, B.W. (2017). Routine activity theory and cybercrime: A theoretical appraisal and literature review. In *Technocrime and criminological theory* (pp. 35–54). Routledge.

Reyns, Bradford W. (2013). Online Routines and Identity Theft Victimization: Further Expanding Routine Activity Theory beyond Direct-Contact Offenses. *Journal of Research in Crime and Delinquency*, *50*(2), 216–238. https://doi.org/10.1177/0022427811425539

Reyns, Bradford W., Henson, B., & Fisher, B. S. (2011). Being pursued online: Applying cyberlifestyle-routine activities theory to cyberstalking victimization. *Criminal Justice and Behavior*, *38*(11), 1149–1169. https://doi.org/10.1177/0093854811421448

Rice, K. J., & Smith, W. R. (2002). Socioecological models of automotive theft: Integrating routine activity and social disorganization approaches. *Journal of Research in Crime and Delinquency*, *39*(3), 304–336. https://doi.org/10.1177/002242780203900303

Robinson, D. (2017). The Incredible Growth of Python. Retrieved from https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rodríguez Manzanera, L. (2007). *Criminología* (22nd ed.). México, D.F.: Editorial Porrúa.

Romeo-Casabona, C. M. (2006). Los datos de carácter personal como bienes jurídicos penalmente protegidos. In *El cibercrimen: nuevos retos jurídico-penales, nuevas respuestas político-criminales* (pp. 167–190). Comares.

Rosen, A., & Ihara, I. (2017). Giving you more characters to express yourself. Retrieved from https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017).

*Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*. https://doi.org/10.17185/duepublico/42132

Ruiz-Funes García, M. (1929). *Endocrinología y criminalidad*. Madrid: Javier Morata.

Sampson, R., Eck, J. E., & Dunham, J. (2010). Super controllers and crime prevention: Au routine activity explanation of crime prevention success and failure. *Security Journal*, *23*, 37–51.

Sánchez, M. A., & Pinochet-Sánchez, G. (2017). El rol de las redes sociales virtuales en la difusión de información y conocimiento: estudio de casos. *Universidad y Empresa*, *19*(32), 107–135. https://doi.org/10.12804/http://revistas.urosario.edu.co/index.php/empresa/article/view/4847

Saunders, K. W. (2011). Hate Speech in the Schools : A Potential Change in Direction. *Maine Law Review*, *64*(1).

Schachaf, P., & Hara, N. (2010). Beyond vandalism: Trolls in Wikipedia. *Journal of Information Science*, *36*(3), 357–370.

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing*. (2012), 1–10. https://doi.org/10.18653/v1/w17-1101

Schneider, K. (1997). *Las personalidades psicopáticas*. Madrid: Triacastela.

Schwartz, M. (2008). The trolls among us. *The Times Magazine*.

Seelig, E. (1958). *Tratado de Criminología*. Madrid: Instituto de Estudios Políticos.

Serrano Maíllo, A. (2006). *Introducción a la Criminología* (4th ed.). Madrid: Dykinson.

Serrano Maíllo, A. (2017). *Teoría criminológica: la explicación del delito en la sociedad contemporánea*. Madrid: Dykinson.

Sharma, S., Agrawal, S., & Shrivastava, M. (2018). *Degree based Classification of Harmful Speech using Twitter Data*. Retrieved from http://arxiv.org/abs/1806.04197

Sheldon, W. (1942). *The varieties of Temperament*. London: Macmillan Pub Co.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, 687–690. Retrieved from http://arxiv.org/abs/1603.07709

Sood, S. O., & Churchill, J. E. F. (2012). Using crowdsourcing to improve profanity detection. *AAAI Spring Symposium - Technical Report*, *SS-12-06*, 69–74.

Spivack, N. (2011). "Web 3.0: The Third Generation Web is Coming". Retrieved from http://lifeboat.com/ex/web.3.0

Sternberg, R. J. (2003). A Duplex Theory of Hate: Development and Application to Terrorism, Massacres, and Genocide. *Review of General Psychology*, *7*(3), 299–328. https://doi.org/10.1037/1089-2680.7.3.299

Suárez, I. (2013). *El Gobierno de Internet España*. España: ISuárez.

Subrahmanyam, K., Reich, S. M., Waechter, N., & Espinoza, G. (2008). Online and offline social networks: Use of social networking sites by emerging adults. *Journal of Applied Developmental Psychology*, *29*(6), 420–433. https://doi.org/10.1016/j.appdev.2008.07.003

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, 177–184.

https://doi.org/10.1109/SocialCom.2010.33

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology and Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sutherland, E. H. (1947). *Principles of Criminology* (4th ed.). Chicago, IL: B.

Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *Delinquency and Drift Revisited: The Criminology of David Matza and Beyond*, *22*(6), 33–41. https://doi.org/10.4324/9781315157962

Taylor, L., Walton, P., & Young, J. (1973). *The new criminology: for a theory of social deviance*. London: Routledge.

Thomasson, E. (2017). German cabinet agrees to fine social media over hate speech. Retrieved from Reuters website: http://uk.reuters.com/article/idUKKBN1771FK.

Thompsen, P., & Ahn, D.-K. (1992). To Be or Not To Be: An Exploration of E-Prime, Copula Deletion and Flaming in Electronic Mail. *A Review of General Semantics*, *49*(2), 146–164.

Timofeeva, Y. A. (2003). Racism versus Freedom of Expression: Current and future approaches to the regulation of Internet hate speech. *Language*, (20).

Tiven, L. (2003). *Hate on the Internet: A response guide for educators and families*.

Tsesis, A. (2001). Hate in Cyberspace: Regulating Hate Speech On the Internet. *San Diego L. Rev.*, *38*, 817.

Túffery, S. (2011). *Data mining and statistics for decision making*. Wiley.

Tulkens, F. (2013). *The hate factor in Political Speech. Where do responsibilities lie*.

Tynes, B. (2006). Children, adolescents and the culture of online hate. In *Handbook of children, culture and violence* (pp. 267–289).

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., … Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. *International Conference Recent Advances in Natural Language Processing, RANLP*, 672–680.

Vozmediano, L., & San Juan, C. (2010). *Criminología ambiental. Ecología del delito y de la seguridad*. Barcelona: UOC.

Vozmediano Sanz, L., San Juan Guillén, C., & Vergara Iraeta, A. I. (2009). Miedo al delito en contextos digitales : un estudio con población urbana. *Eguzkilore*, *23*, 175–190.

Waldron, J. (2012). *The harm in hate speech*. London: Harvard University Press.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, (Lsm), 19–26. Retrieved from http://dl.acm.org/citation.cfm?id=2390374.2390377

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. https://doi.org/10.18653/v1/W16-5618

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Wasilewska, A., & Menasalvas, E. (2008). Data Preprocessing and Data Mining as Generalization. In *Data Mining: Foundations and Practice* (pp. 469–484). Springer Berlin Heidelberg.

Weber, A. (2009). *Manual on hate speech*. Council of Europe.

Wei, X., Lin, H., Yang, L., & Yu, Y. (2017). A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information (Switzerland)*, *8*(3), 1–16. https://doi.org/10.3390/info8030092

Wendling, M. (2015). *2015: The year that angry won the internet*. Retrieved from https://www.bbc.com/news/blogs-trending-35111707

Wilcox, R. . (1998). *Wings of Fury: True Story of America's Elite Fighter Pilots*. London: Simon & Schuster Ltd.

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Xue, M., Yang, L., Ross, K. W., & Qian, H. (2017). Characterizing user behaviors in location-based find-and-flirt services: Anonymity and demographics: A WeChat Case Study. *Peer-to-Peer Networking and Applications*, *10*(2), 357–367. https://doi.org/10.1007/s12083-016-0444-5

Yar, M. (2005). The Novelty of 'Cybercrime': An Assessment in Light of Routine Activity Theory. *European Journal of Criminology*, *2*(4), 407–427. https://doi.org/10.1177/147737080556056

Ybarra, M. L., Diener-West, M., Markow, D., Leaf, P. J., Hamburger, M., & Boxer, P. (2008). Linkages between internet and other media violence with seriously violent behavior by youth. *Pediatrics*, *122*(5), 929–937. https://doi.org/10.1542/peds.2007-3377

Yuan, S., Wu, X., & Xiang, Y. (2016). A Two Phase Deep Learning Model for Identifying Discrimination from Tweets. *EDBT*, 696–697. Retrieved from http://arxiv.org/abs/1308.0850

Yucedal, B. (2010). *Victimization in Cyberspace: An Application of Routine Activity and Lifestyle Exposure Theories*. Kent State University.

Zarras, A., Kapravelos, A., Stringhini, G., Holz, T., Kruegel, C., & Vigna, G. (2014). The dark alleys of madison avenue: Understanding malicious advertisements. *Proceedings of the 2014 Conference on Internet Measurement Conference*, 373–380. https://doi.org/10.1145/2663716.2663719

Zhang, Z., Robinson, D., & Tepper, J. (2018). Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network. *Eurpoean Semantic Web Conference*, 745–760. Retrieved from https://doi.org/10.475/123_4

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5

# ANEXOS

## ANEXO I. CÓDIGO FUENTE DE LA HERRAMIENTA INFORMÁTICA

```
import pandas as pd
import numpy as np
import re
import codecs
import time
from datetime import datetime
import pydotplus
import random
from sklearn.ensemble import RandomForestClassifier
from sklearn.cross_validation import cross_val_score
from sklearn import metrics
from sklearn import tree
#
================================================================
=================
# Read XLSX file
#
================================================================
=================
xl = pd.ExcelFile("BBDD/LondonDefinitiva_clasif200880.xlsx")
data = xl.parse("Sheet1")
#
================================================================
=================
# Select important variables
#
================================================================
=================
data = data[["text", "created_at", "retweet_count", "description", "geo_enabled",
        "user_created_at", "listed_count", "statuses_count", "followers_count",
        "friends_count", "favourites_count", "verified", "Do_Dv"]]
```

```
#
================================================================
================
# Extract new variables
#
================================================================
================
txt_twt = data["text"] #Tweet text

# Extract "mention_count", "hashtag_count", "emoji" and "text_count"
vector_mention = []
vector_hashtags = []

#Find  URL and remove URL y emoji of tweet text
stopword = []

cont_url = 0
cont_emoji = 0 #Emoji sí o no

vector_url = []
vector_emoji = []

#Contadores bucle
i=0
x=0

for i in range(len(txt_twt)): #Recorre los tweets de la muestra

    #---------------- Eliminar "RT @usuario: " ---------------------------------
   if (txt_twt[i][0:4] == "RT @"):
      #stopword.append([" ".join(txt_twt[i].split(" ",2)[0:2])])
      l_rtwt = len(" ".join(txt_twt[i].split(" ",2)[0:2]))+1
      data.loc[i, 'text'] = txt_twt[i][l_rtwt:]
```

```
#------------------------- Inicializar contadores ----------------------
cont_mention = 0 #Contador menciones
cont_hashtags = 0 #Contador hashtags
cont_url = 0
cont_emoji = 0

#---------------------- Busca e introduce en un array las URLs ------------
url = txt_twt[i].find("http")

if (url > -1): #Si es -1 no hay Url
    stopword.append([ t for t in txt_twt[i].split() if t.startswith('http') ])
    #Si encuentra URL lo introduce en el array
    cont_url =  1
#Introduce en el vector si hay o no URL en el texto del tweet
vector_url.append(cont_url)

#------------------ Busca e introduce en un array los emojis --------------
emojis = txt_twt[i].find("<")

if (emojis > -1): #Si es -1 no hay Url
    stopword.append([ t for t in txt_twt[i].split() if t.startswith('<') ])
    cont_emoji = 1

#Introduce en el vector si hay o no EMOJI en el texto del tweet
vector_emoji.append(cont_emoji)



#---------------- Contar hashtags y menciones -----------------------------
for x in txt_twt[i]:#Recorre caracteres del text
    #Mentions o hashtags
    if(x=="@"):
```

```
            cont_mention = cont_mention + 1
        if(x=="#"):
            cont_hashtags = cont_hashtags + 1

    #Añadir contador a los vectores
    vector_mention.append(cont_mention)
    vector_hashtags.append(cont_hashtags)

    x = 0 #Reiniciamos el conteo de caracteres

#Introducir el conteo en la base de datos original
data["mention_count"] = vector_mention
data["hashtag_count"] = vector_hashtags
data["url"] = vector_url
data["emoji"] = vector_emoji

#
================================================================
=================
# DATE TWEET
#
================================================================
=================
created_at = data["created_at"]
vector_created_at = []

i=0
for i in range(len(created_at)):
    time_struct = time.strptime(created_at[i], "%a %b %d %H:%M:%S +0000 %Y")
#Tue Apr 26 08:57:55 +0000 2011
    date = str(datetime.fromtimestamp(time.mktime(time_struct)))
    vector_created_at.append(date)
```

```
data['created_at'] = vector_created_at

#Difference between two date
start_date = datetime.strptime("2017-06-03 23:00:00","%Y-%m-%d %H:%M:%S")

end_date = data["created_at"]
vector_minute = []

i=0
for i in range(len(end_date)):
    end = datetime.strptime(end_date[i],"%Y-%m-%d %H:%M:%S")

    difference = end - start_date
    minutes = int(difference.total_seconds() / 60)

    vector_minute.append(minutes)

data["minute_count"] = vector_minute

#
================================================================
================
# Write XLSX
#
================================================================
================
directory = "/Users/crimina/Documents/Investigaciones/Papers/Hate is in the
air/scripts"

writer = pd.ExcelWriter(directory + '/output.xlsx')
data.to_excel(writer,'output')
writer.save()
```

```
#
================================================================
================
# # MODELO
#
================================================================
================

model = RandomForestClassifier(
 random_state      = 1,  # semilla inicial de aleatoriedad del algoritmo
 n_estimators      = 1000, # cantidad de arboles a crear
 max_depth         = 10,
 n_jobs            = -1    # tareas en paralelo. para todos los cores disponibles usar -1
 )
model.fit(X = data_train_analysis, y = clase_train)



# PREDICCION
#-------------------------------------------------------------------------------------------
prediction = model.predict(data_test_analysis)



# METRICAS
#-------------------------------------------------------------------------------------------
print(metrics.classification_report(y_true=clase_test, y_pred=prediction))
print(pd.crosstab(clase_test, prediction, rownames=['REAL'],
colnames=['PREDICTION']))

# IMPORTANCIA VARIABLES
#-------------------------------------------------------------------------------------------
var_imp = pd.DataFrame({
 'feature':headers,
 'v_importance':model.feature_importances_.tolist()
```

```
 })
print (var_imp.sort_values(by = 'v_importance', ascending=False))
```

ANEXO II. INDICIO DE CALIDAD DE LA TESIS DOCTORAL (ARTÍCULO CIENTÍFICO)

# CAN METADATA BE USED TO MEASURE THE ANONYMITY OF TWITTER USERS? RESULTS OF A CONFIRMATORY FACTOR ANALYSIS[6]


## Authorship

Zoraida Esteve Bañón, Crímina Research Centre for the Study and Prevention of Crime, Miguel Hernandez University of Elche.

Asier Moneva, Crímina Research Centre for the Study and Prevention of Crime, Miguel Hernandez University of Elche.

Fernando Miró-Llinares, Crímina Research Centre for the Study and Prevention of Crime, Miguel Hernandez University of Elche.



## Abstract

Anonymity is one of the elements traditionally associated with criminal and antisocial behaviour. Anonymity depends on several factors, such as natural surveillance or the visibility created by the physical or digital environment. Certain digital environments, such as social networks, exhibit characteristics that facilitate or limit the degree of anonymity of their users. Social networks are places in cyberspace



[6] Este estudio ha recibido el apoyo del Instituto Nacional de Ciberseguridad (INCIBE) en el marco de las "Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad" (ref. INCIBEI-2015-27349). Artículo recuperado de https://www.ehu.eus/ojs/index.php/inecs/article/view/21157 Autor de correspondencia: Universidad Miguel Hernández de Elche. Avda. de la Universidad s/n. Edif. Hélike (CRIMINA), 03202, Elche (España). (+34) 965 22 20 85, z.esteve@crimina.es

where people interact with each other and with the environment, where they increasingly carry out their daily activities and where they also commit crimes. This paper attempts to test the hypothesis that certain elements of the social network environment define the anonymity of their users. To this end, an empirical process for quantifying anonymity is proposed, which can be applied transversally to all places in cyberspace that permit user accounts. Subsequently, a data set of 162 users has been obtained from the social network Twitter which also collects the metadata associated to their accounts. To test this hypothesis, a Confirmatory Factor Analysis (CFA) has been conducted to determine whether the data obtained fit the model based on a theoretical concept proposed by the researchers. The results show a moderate fit for the model, suggesting that some metadata (i.e., geopositioning) do not contribute to defining the latent variable anonymity. We suggest the proposed model needs to be reconsidered and applied to a larger sample to improve its fit. Finally, the applicability of the proposed methodology for measuring anonymity and future lines of research are discussed.

**Keywords**

Anonymity, cyberspace, metadata, Twitter, CFA.

## 1. Introducción

Information and Communication Technologies (ICT) have revolutionized the way we relate to each other. First the growth of the Internet and later the growth of ICT as a whole, have been fundamental in creating an environment which is distinct from physical space and which allows content sharing and instant interaction with other members of the virtual community. Thus, cyberspace, understood as a virtual space which resembles a geographic space without being the same (Miró-Llinares & Johnson,

2018), has altered the meaning of distance and time, which are now compressed to the point of disappearance (Miró-Llinares, 2011, 2012). Along with these intrinsic characteristics, cyberspace has been configured by other extrinsic characteristics such as its transnationality, neutrality, decentralization, universality and anonymity, attributes that have driven its popularization, but that have also imposed obstacles to the prevention and prosecution of the crimes that occur within it (Miró-Llinares, 2011). Thus, for example, the absence of barriers, which in the physical space represent the borders that configure the different states, makes it enormously difficult for the justice system to act outside its boundaries, lost in a labyrinth of infinite and complex legal rules and procedures. The neutral and non-centralized nature of the Internet reduces or practically eliminates restrictions on accessing websites or disseminating information, making it difficult to control the flow of content and, therefore, also the behaviour of its users. Finally, anonymity, the driving force behind the popularisation of the Internet, inhibits social controls (Armstrong & Forde, 2003; Finn, 2004; McGrath & Casey, 2002) and provides criminals with an attractive environment (McGrath & Casey, 2002) by increasing the sense of impunity for crime and dissociating their anonymous online actions from their behaviour in physical space (Suler, 2004).

In this sense, scientific literature has extensively described the role of anonymity in the genesis of criminal behaviour, both in physical space (Haney, Banks, & Zimbardo, 1973; Lelkes, Krosnick, Marx, Judd, & Park, 2012; Rogers & Ketchen, 1979), and in cyberspace (Baggili & Rogers, 2009; Hinduja, 2008; Ševčíková & Šmahel, 2009). Thus, research such as that of Armstrong and Forde (2003) on paedophilia or that of Finn (2004) on harassment have shown that online environments can favour a false sense of intimacy, as well as uninhibited behaviour that increases risk-taking and antisocial behaviour. It seems, therefore, that it is definitely the virtual space which, given its

specific configuration favouring anonymous behaviour, provides the appropriate conditions for criminal events to occur. Thus, by focusing on this new place we have called cyberspace, we find ourselves required to ask whether it makes sense to study crime from a theoretical perspective, under the same postulates as we do in physical space. Hence, when we examine the consistency of the so-called crime theories applied to this "no place" (Miró-Llinares & Johnson, 2018) we find that, if the minimum elements present in the criminal event are the same, that is, a potential offender, a suitable target, and the absence of a capable guardian, whose convergence occurs in a (digital) place and at a specific moment (Cohen & Felson, 1979; Miró Llinares, 2011), it seems logical to think that the fundamental premises of crime theories can be adapted to help explain the commission of crimes in cyberspace.

In any case, we know that people's daily activities are moving progressively into cyberspace and that, as in physical space, criminal opportunities are not randomly distributed in cyberspace, but are concentrated in certain places where the risk of a crime occurring is greater (Miró-Llinares & Johnson, 2018). In this way, users who buy on certain pages and do not check their security before making payment become appropriate targets for cyber-scammers who are aware of such vulnerability (Pratt, Holtfreter, & Reisig, 2010). Similarly, users who use email to exchange messages and files with others are more likely to receive spam (Yeargain, Settoon, & McKay, 2004) or suffer a malware infection (Hoar, 2005). It is true that in all these crimes and in others such as hate speech, the convergence between the aggressor and the victim takes place in a different way from traditional offences. However, convergence is still necessary, as it is essential that there is a place where the hate message is expressed and where another user receives it (Miró-Llinares, Moneva, & Esteve, 2018). Thus, just as with certain crimes in physical space, such as theft, which tend to be concentrated in the

places where businesses are located (Wikström, 1995), in cyberspace both crime and the perception of insecurity are also distributed according to the characteristics of cyberplaces (Castro-Toledo & Miró-Llinares, 2018; Miró-Llinares & Johnson, 2018; Miró-Llinares et al, 2018) and the interaction between users, such as forums, chats and, mainly, social networks.

The popularization and universality of social networks means that virtual convergence among users is sometimes more frequent than in physical space, which in turn increases criminal opportunities. For example, on the social network Twitter, many users publish real personal information such as multimedia content, location, routines, etc. that put them in a position of vulnerability with regards to certain cybercriminals. On the contrary, other users interact anonymously, hiding their identity through a pseudonym or false names (Peddinti, Ross, & Cappos, 2014), which allows them to express opinions and publish sensitive information without fear of being identified (Peddinti, Ross, & Cappos, 2017b). Although not all Twitter users hide behind a veil of anonymity to commit criminal or deviant behaviour, the work of Peddinti and colleagues (2017b; 2017a) shows that there is a relationship between the anonymity of users and the pornographic, homophobic and Islamophobic content published from their accounts. Both the anonymity provided by this social network and the ease with which it is accessed, which allows users with different personal characteristics to identify with certain ideologies (Perry & Olsson, 2009), have favoured Twitter becoming a platform where some users emit radical and hateful messages that remain fixed over time, thereby reaching a massive audience and thus increasing their harmfulness (Miró-Llinares et al., 2018). However, scientific research has not yet shown what the specific influence of anonymity is on the commission of criminal behaviour on Twitter, mainly due to the methodological difficulties involved in quantifying this condition.

## 2. A step-by-step proposal to measure the anonymity of online users

There is some consensus on the existence of a relationship between anonymity on the Internet and criminal behaviour, despite the fact that few empirical studies have delved into its influence at the individual level and despite the fact that some who have done so tangentially have not extracted conclusive results (Bautista-Ortuño, 2017). Perhaps one of the reasons for this derives from the fallacious understanding of cybercrime as a single event when, in reality, there are multiple criminal modalities of very different nature and in which anonymity can play a very different role (Miró-Llinares, 2015). In addition, cyberspace is not univocal either, so it seems unrealistic to try to measure all the factors that configure "anonymity on the Internet", since they vary according to the configuration of each digital environment. In this sense, the few exceptions that have attempted to define and measure anonymity in cyberspace rightly try to circumscribe it to certain places. Peddinti et al. have done so on Twitter via an approach which is more similar to analysis of deviance than crime (Peddinti, Korolova, Bursztein, & Sampemane, 2014; Peddinti, Ross, & Cappos, 2017b, 2017a, 2014) and others have tried to replicate their methodology in isolation (Xue, Yang, Ross, & Qian, 2017). According to these investigations, anonymous Twitter users are more uninhibited, interact more, follow more accounts and are more willing to display their activity to the general public. The scale of anonymity developed by Peddinti and colleagues is, however, debatable as a method for classifying Twitter users according to their anonymity, especially with a view to analysing their relevance in relation to criminal or antisocial activity. The authors divide Twitter users into four categories according to their degree of anonymity:

- Anonymous: The user has not provided a first and last name or a URL in their profile.

- Partially anonymous: The account contains a first name or last name, but not both in their profile.
- Identifiable: The user has indicated their name and surname in their profile.
- Unclassifiable: The user does not indicate name or surname, but they do have a URL in their profile (p.85).

Although this is an interesting first approximation, identifying anonymity exclusively with the wording of a name or surname, or with the presence of a URL in the profile, does not seem to be an adequate strategy to measure the degree of anonymity with which a user acts at a specific time. It is even less so if we bear in mind that Twitter is characterized by intense visualization of contents that are published from very heterogeneous profiles. However, given the different structural and communicative nature of the different environments that make up the Internet, we do think it is opportune to confine this analysis to a specific place in cyberspace.

This paper proposes a methodology to measure the degree of anonymity of online users by following a sequential and systematic five-phase process (Figure 1): (1) select a specific cyber place; (2) identify the relevant metadata from a user's profile; (3) collect additional relevant information associated with the profile; (4) operationalize the identified anonymity variables; and (5) conduct a Confirmatory Factor Analysis (CFA) to model the anonymity of users.

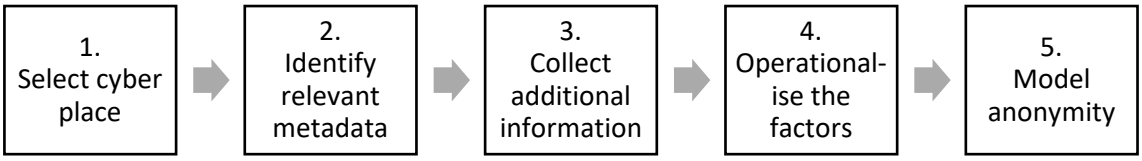


Figure 1. Systematic five-step sequential process to measure online users' anonymity

### *2.1. Selection of a specific cyber place*

Since each cyber place has a different configuration that conditions the degree of exposure of its users, the first step to measure exposure is to select a specific digital environment that will be subject to the analysis. According to Miró-Llinares and Johnson (2018), cyber places can be classified according to (1) the type of contact they allow, (2) the natural surveillance they enable and the self-protection measures they make available to their users, and (3) the type of activity carried out there. All cyber places are configured to a certain extent by the way in which these three elements are combined. Therefore, it is essential to pay attention to these characteristics and how they relate to the potential anonymity of the users who visit them.

Regarding the former, most cyber places have store-and-forward information transmission channels, although some of them, such as the social network Twitter, also include streaming contact systems such as Periscope. The fundamental implication in relation to anonymity is that while the first modality implies greater control for the user over the information that they wish to transmit and the way to do it, since they can dedicate an indeterminate time to plan its publication, the second requires greater capacity for improvisation and, therefore, risks are assumed. Although the essential form of communication on Twitter is the tweet (i.e., any message published on Twitter that can contain photos, videos, links and text), it is possible to communicate privately through direct messages. The second characteristic that defines cyber places is their natural surveillance, defined by the volume of information traffic generated by their users and the degree of publicity of the content they include. In this sense, Twitter is a social network with a significant daily influx of users and which offers users a series of self-protection tools that allow filtering of certain messages or blocking annoying content. In addition, the configuration of social networks in general, and Twitter in

particular, allows the user's privacy to be adjusted. With respect to the third element, the configuration of cyber places determines to a great extent the routines of its users and, therefore, the type of activity carried out there. While some users use Twitter for professional purposes - providing their real identity, establishing a network of professional contacts, and paying special attention to the content they publish from their profile (for example, academic) - others do not need to provide personal information, as they use the social network for leisure.

In addition, cyber places can be analysed at the macro or micro level. Once you have selected the macro-place where you want to measure the anonymity of your users, the next step is to identify the micro-place that contains the relevant information related to them. In the macro cyber place "Twitter" this information can be found in the micro cyber place "user account" (Miró-Llinares et al., 2018).

### *2.2. Identification of the relevant metadata associated with the user profile*

In addition to the personal information collected, these accounts generate additional information linked to each user that reflects their interaction and, in turn, defines their public exposure. For example, a Twitter user's account contains information related to both visibility and anonymity (Miró-Llinares et al., 2018). Regarding the second factor, each Twitter account contains at least the following information: (1) if it is a verified account, (2) the user's name, (3) if it adds a short biography, (4) if it is a geopositioned account, (5) an external link, (6) if it includes a location, (7) if it has a profile photo, (8) if it has a cover photo, and (9) if it has other photos. The first of these elements has definite importance for determining the degree of anonymity of a user, since if Twitter grants a user the blue badge, it means they have passed a rigorous process of identity verification. Given that the rest of the information can be falsified to a greater or lesser

extent, a subsequent categorization is necessary to reflect this and to establish a range depending on the greater or lesser anonymity provided by each element.

### *2.3. Collect information from the profile*

Although some of the anonymity-related factors are only accessible through the Application Programming Interface (API) that Twitter only makes available to users with developer permissions, others are publicly accessible and can therefore be collected manually. In relation to the API, access to this information is increasingly restricted, as recent problems related to the filtering of users' personal information from the social network Facebook has affected privacy and has called into question the ability of these companies to ensure the protection of user data. Nevertheless, provided that the applicant meets a series of requirements and, in some cases, passes an evaluation regarding the justification it has provided for access to such information, a significant number of APIs can be accessed that in turn facilitate access to information stored on social network servers such as Twitter, Facebook, or YouTube, among others. In the case of manual collection processes, it is necessary to adopt a strategy for the systematic observation of the cyber places described in phase 1 that gather the information identified in phase 2. However, when accessing each of the user profiles to record the appropriate information, difficulties may arise related to the privacy settings that each user has set up for their profile, or the account may have been sanctioned by Twitter and closed accordingly. In any case, the identification of the relevant metadata in order to capture the degree of exposure of an online user is not only a characteristic of social networks but, rather, it is transversal to all those cyber places configured to store user profiles.

### *2.4. Operationalization of anonymity factors*

Once the relevant variables have been selected, the next step is to proceed with their operationalization. The approach to operationalising these variables depends on the analysis technique that is used afterwards, so there is no single criterion to carry out this process. However, it is preferable to opt for a quantitative categorization whenever possible, as these data are more flexible to treatment and allow the variables to be re-operationalised in a qualitative format if necessary. On the other hand, a qualitative operationalisation of the variables will not allow the inverse step to be carried out with such a high level of detail. However, it is true that it is sometimes complex to quantify certain characteristics and a qualitative operation must be used, in which case it is essential that the resulting variable is ordinal. The values taken by the variables must describe a range that goes through an anonymity-exposure scale in order to be able to interpret the results of the analyses in one sense or another; that is, it is necessary to know whether a category corresponds to a greater or lesser degree of anonymity in order to determine the direction of the resulting relationship. For example, it would be possible to quantify the number of photos that a user has posted from their account or it would be possible to discretise this information to know if the user has ever posted a photo or not. While from the first method it is always possible to obtain the coding for the second, knowing whether photos have ever been published from a profile does not allow us to know to what extent this is so. Other metadata such as name or location do not allow quantification, so they should be used as a nominal dichotomous scale. In this case, the absence of the attribute is considered an increase in the anonymity of the user and its presence the opposite.

### *2.5. Modelling anonymity*

The analytical strategy adopted to measure anonymity depends on how the data have been recorded and categorized. In this pilot study, an approach to the anonymity model is proposed through the application of a Confirmatory Factor Analysis (CFA). CFA is a type of factorial analysis that is used to model a latent variable whose value is unknown with a set of manifest variables whose value is known. A fundamental requirement for the application of CFA is that the relationships established between the manifest and hypothetical variables within the proposed model are guided by a solid theoretical approach (Schreiber, Nora, Stage, Barlow, & King, 2006). In this case, the variables based on metadata identified in phase 2 have been included in a model to adjust the latent variable that reflects the degree of anonymity-exposure of each user and that has been defined theoretically from the previous variables. In order to execute this statistical technique, the free software R is used, loading the functionalities offered by the 'lavaan' package (Rosseel, 2012).

## 3. The pilot study

The purpose of this study is to illustrate an application of the proposal to measure the anonymity of online users by executing a CFA on a sample of Twitter users. The aim is to determine whether these data fit the anonymity model previously hypothesized by the researchers.

### *3.1. Sample*

After acquiring developer permissions on the Twitter platform, it is possible to access the Streaming API to make a formal request for information about the online activity of users of this social network. After the terrorist attack on London Bridge in June 2017, a request was made to the Twitter server to obtain a sample of the tweets

issued by users in order to study the prevalence of violent communication and hate speech in this environment of digital interaction. In order to delimit the query, a filtering criterion by language and keywords was established. Regarding the first criterion, only those messages published in English were requested. Regarding the second criterion, three keywords were defined based on the hashtags that at the time of the request were global trending topics on Twitter: #LondonBridge, which sought to identify those messages referring to the terrorist event from a neutral position; #PrayForLondon, which sought to filter messages of solidarity or support; and #StopIslam, which sought to collect expressions of negative or discriminatory content. This query returned a sample of 200,882 records in the unstructured JSON format containing information associated with Twitter users and the tweets they had published after the attack (Miró-Llinares et al., 2018). After a process of classifying the messages with a criterion of inter-rater agreement trained in the Taxonomy of hate and violent communication on the Internet (Miró-Llinares, 2016), an additional dichotomous attribute was assigned to each record indicating whether the message could be included within the categories defined in the taxonomy or not.

To carry out the proposed pilot study, 200 users were selected who had published a message in the sample described; 100 of whom had published at least one hate message. Subsequently, each of the selected profiles was accessed to check whether the user account is still active. Given the impossibility of collecting some essential data to conduct the CFA model proposed here (i.e., suspended or closed accounts), 38 of these users were excluded from the sample. The final number of users included in the sample for the systematic observation of their profile is 162. Since this study is going to apply a CFA on a single sample and taking into account the consensus established in the specialized literature that suggests the incorporation of 10 records for each of the

estimated parameters is acceptable, the sample size is adequate (for example, Schreiber et al., 2006).

### *3.2. Manifest variables*

After excluding the variable that determines whether a user account is verified, the 8 remaining variables identified in phase 2 were selected and the data collected following a process of systematic observation of the 162 identified profiles. Subsequently, the variables were quantitatively categorised by describing an anonymity-exposure range, as indicated in phase 3. Table 1 below summarizes the characteristics of the manifest variables included in the model.

Table 1. Operationalisation of the manifest variables for the CFA

| Variable | Categorisation |
|---|---|
| Name | 0: the user name is fictitious; 1: the user name can be real |
| Biography | 0: the user does not have a biography; 1: the user's biography does not provide relevant information about the user's identity; 2: the user's biography provides information that may be relevant to identify the user. |
| Geopositioning | 0: the user has not activated the geopositioning of their tweets; 1: the user has activated the geopositioning of their tweets. |
| External URL | 0: the user does not include a URL in their profile; 1: the website to which the profile URL redirects does not provide relevant information about their identity; 2: the website to which the profile URL redirects may be relevant to identify them. |
| Location | 0: the user does not include a location in their profile; 1: the user includes a location in their profile that is fictitious; 2: the user includes a location in their profile that may be real |
| Profile photo | 0: user does not include a profile picture; 1: user includes a profile picture that does not show a person; 2: user includes a profile picture of a person |
| Cover photo | 0: the user does not include a cover photo; 1: the user includes a cover photo that does not show a person; 2: the user includes a cover photo that shows a person. |
| Other photos | 0: the user has not posted messages with images; 1: the user has posted messages with images that do not show people; 2: the user has posted messages with images that show a person. |

For the CFA model, we conducted 1 regression for each of the manifest variables, giving a total of 8. Taking into account the nature of the variables included in the model, the use of a robust weighted least squares estimator (WLSMV; Rosseel, 2012) has been selected.

## 4. Results

Since our model assumes the existence of relationships between the variables that define it, the standardized covariance matrix of the variables has been extracted to observe the differences between the expected and observed correlations in the model (Table 2). Values > 0.1 are indicators of relationships that can be conflicting when adjusting the model and that will be reflected in the error measurement. In general, these residual correlations show acceptable values, with some exceptions (for example, name, geopositioning, location).

Table 2. Covariance matrix of variables for CFA

| Manifest variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. Name | | | | | | | | |
| 2. Biography | -0.12 | | | | | | | |
| 3. Geopositioning | 0.12 | -0.13 | | | | | | |
| 4. External URL | -0.01 | 0.10 | -0.03 | | | | | |
| 5. Location | 0.07 | 0.07 | 0.16 | 0.14 | | | | |
| 6. Profile photo | 0.16 | 0.01 | 0.09 | -0.09 | -0.09 | | | |
| 7. Cover photo | -0.08 | -0.03 | -0.07 | -0.10 | -0.17 | 0.03 | | |
| 8. Other photos | -0.12 | -0.09 | -0.08 | 0.00 | -0.06 | -0.05 | 0.09 | |

Note. The variables have been standardized in accordance with the following parameters: $M$ = 0.00; $SD$ = 1.00 ($N$ = 162).

Below are the main indicators of the fit of the specified model based on the main indices that, according to Schreiber and colleagues (2006; see also Kline, 2011), are

fundamental in evaluating unique analyses such as the one presented in this study: (1) the Tucker-Lewis index (TLI), which should be ≥ 0.96 for categorical data for a good fit; (2) the comparative fit index (CFI), which should be ≥ 0.95; and (3) the root mean square error of approximation (RMSEA), which should be < 0.60. While the first two serve to compare the model, the latter is a measure of error based on the residual correlations reflected in Table X. In our model the indices described take the following values: TLI = 0.98; CFI = 0.99; RMSEA = 0.05. In its robust version, the values are: TLI = 0.98; CFI = 0.99; RMSEA = 0.05. In its robust version, the values are: TLI = 0.98; CFI = 0.99; RMSEA = 0.05: TLI robust = 0.95; CFI robust = 0.96; RMSEA robust = 0.07. These results show a moderate degree of fit between the model and the observed data, since, although normal indices give good results, their robustness does not exceed the cut criteria in two of the three cases.

As expected, all variables included in the model have factor loads that are significantly related to the latent anonymity construct. Most standardized coefficients (i.e., Betas) are above the threshold of 0.60, although it is true that they range from 0.26 regarding geopositioning to 0.74 with regards to other photos (Table 3). Figure 2 shows the proposed CFA model for the anonymity latent variable with standardized parameters for each manifest variable. Although the coefficients associated with the geopositioning variable suggest that discarding it could improve the fit of the model, no post hoc modifications have been made, since the hypothesized model derives from a sufficiently justified theoretical approach.

*Table 77. Regression coefficients for the variables in the CFA*

| Manifest Variable | B | SE | Z | Beta | sig. |
|---|---|---|---|---|---|
| Name | 1.00 | 0.09 | 4.95 | 0.43 | *** |
| Biography | 1.58 | 0.07 | 10.05 | 0.68 | *** |
| Geopositioning | 0.60 | 0.11 | 2.44 | 0.26 | * |
| External URL | 1.58 | 0.06 | 11.41 | 0.68 | *** |
| Location | 1.53 | 0.07 | 8.99 | 0.66 | *** |
| Profile photo | 1.42 | 0.06 | 9.63 | 0.62 | *** |
| Cover photo | 1.60 | 0.07 | 9.96 | 0.69 | *** |
| Other photos | 1.72 | 0.06 | 12.88 | 0.74 | *** |

Note. * p value < 0.05; ** p value < 0.01; *** p value <0.001

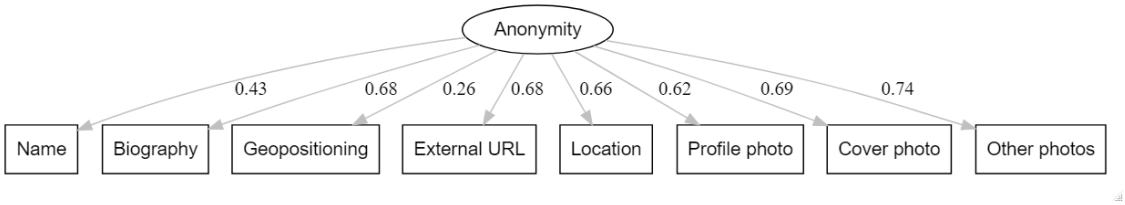


Figure 29. CFA Model for latent variable anonymity. Robust Chi-squared = 0.015. Degrees of freedom = 20.

## 5. Discussion and conclusions

In the present study a CFA has been performed to test the hypothesis that the metadata of Twitter users' accounts (i.e., manifest variables) define their anonymity (i.e., latent variable) measured as a theoretical construct. Since the proposed model is based on both a conceptualization exercise and a review of the literature on the operationalization of anonymity on Twitter (Miró-Llinares et al., 2018; Peddinti, Korolova, et al., 2014; Peddinti, Ross, et al., 2014; Peddinti et al., 2017a, 2017b), a CFA has been chosen rather than an Exploratory Factor Analysis (EFA). Unlike the CFA, the EFA pursues the creation of new latent variables starting from groups of factors whose relationship is unknown from a theoretical point of view. It can be affirmed that while the CFA is guided by theory (Schreiber et al., 2006), the EFA is guided by pragmatics (Tabachnick, Fidell, & Ullman, 2019), and while at the time of executing a CFA the variables of the model are previously defined, the EFA serves precisely to define the number of variables that define a construct (Williams, Onsman, & Brown, 2010). For the same reason, even when the results do not show a good fit for the data, when a CFA is performed it is not necessary

to include or eliminate factors to improve the fit of the model. In the present case, the CFA results show that the proposed manifest variables, modelled as a single linear combination of factors, show a moderate fit; that is, the proposed model can be said to have limited capacity to explain the latent variable anonymity. In summary, the data partially support the proposed hypothesis and show that it is possible to develop more rigorous anonymity measurement models than those currently in use.

Beyond the results, another contribution of this paper is the proposal to quantify anonymity. In contrast to other attempts to measure user anonymity on Twitter (for example, Peddinti et al., 2017b), the proposed process is not only more exhaustive, as it makes use of metadata associated with users' online accounts, which allows for the incorporation of a wide variety of elements that characterise such a digital environment, but also transcends the social network Twitter, as its sequential design is devised for application in any digital environment configured to record and store user accounts. Thus, for example, this proposal can be extrapolated to email user accounts, forums, or web applications. The versatility of this methodology makes it possible to explore the influence of anonymity, as well as other constructs, on criminal behaviour and deviant behaviour in cyberspace.

Regarding the method of data collection, researchers have used the open data policy of Twitter to make a request for information that has allowed access to certain information that otherwise would not be accessible and that has subsequently been supplemented with additional information collected manually. It should be noted that other social networks do not allow access to their data for research purposes or restrict it altogether. The same obstacle may arise in the case of other cyber places that also involve user accounts. Therefore, in order to follow the process of quantification of anonymity proposed in this paper, it will be necessary to rely on manual data collection methods that will only be

valid to the extent that they are also systematic and rigorous. These methods pose a number of problems: (1) some factors that hypothetically allow latent variables to be defined from theoretical approaches will not be accessible; (2) the costly sample collection process will greatly limit the ability to obtain large amounts of information that would allow more robust models to be developed; and (3) such a process is difficult for other researchers to replicate, limiting the ability to contrast results or extrapolate them to other similar contexts for comparative analysis.

Future research on the use of CFA to measure the anonymity of online users should work on improving the process of obtaining and coding the variables included in the model to improve their fit. To this end, it is necessary to identify new variables related to the anonymity of the accounts and to propose new ways of operationalizing the existing ones in order to achieve greater completeness and precision in their modelling. Secondly, it is necessary to carry out studies with a larger sample to give greater robustness to the results obtained, so it is advisable to use the APIs as the main way of obtaining data. Finally, it would be interesting to see to what extent the new construct of anonymity obtained after modelling acts as a predictor of criminal or deviant behaviour in cyberspace (for example, hate speech, spam, fraud), as its relationship with cybercrime is often taken for granted, but its empirical study in cyberspace has been neglected.

## References

Armstrong, H. L., & Forde, P. J. (2003). Internet anonymity practices in computer crime. Information Management & Computer Security, 11(5), 209–215. https://doi.org/10.1108/09685220310500117

Bautista-Ortuño, R. (2017). ¿Eres un ciberhater? Predictores de la comunicación violenta y el discurso del odio en Internet. International E-Journal of Criminal Sciences, 11, 1–28.

Castro-Toledo, F. J., & Miró-Llinares, F. (2018). ¿Nos parecen más inseguros los ciberlugares después de un ciberataque? International E-Journal of Criminal Sciences, 12, 1–25.

Finn, J. (2004). A Survey of Online Harassment at a University Campus. Journal of Interpersonal Violence, 19(4), 468–483. https://doi.org/10.1177/0886260503262083

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed). New York: Guilford Press.

McGrath, M. G., & Casey, E. (2002). Forensic psychiatry and the internet: Practical perspectives on sexual predators and obsessional harassers in cyberspace. The Journal of the American Academy of Psychiatry and the Law, 30(1), 81–94.

Miró-Llinares, F. (2015). Cibercrimen y vida diaria en el mundo 2.0. In F. Miró-Llinares, J. R. Agustina-Sanllehí, J. E. Medina-Sarmiento, & L. Summers (Eds.), Crimen, oportunidad y vida diaria (Dykinson, pp. 415–456). Madrid.

Miró-Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet. IDP: Revista de Internet, Derecho y Política, 22, 82–107.

Miró-Llinares, F., & Johnson, S. D. (2018). Cybercrime and Place: Applying Environmental Criminology to Crimes in Cyberspace. In G. J. N. Bruinsma & S. D. Johnson (Eds.), The Oxford Handbook of Environmental Criminology (pp. 883–906). https://doi.org/10.1093/oxfordhb/9780190279707.013.39

Miró-Llinares, F., Moneva, A., & Esteve, M. (2018). Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Science, 7(15), 1–12. https://doi.org/10.1186/s40163-018-0089-1

Peddinti, S. T., Korolova, A., Bursztein, E., & Sampemane, G. (2014). Cloak and Swagger: Understanding Data Sensitivity through the Lens of User Anonymity. 2014 IEEE Symposium on Security and Privacy, 493–508. https://doi.org/10.1109/SP.2014.38

Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). 'On the internet, nobody knows you're a dog': A twitter case study of anonymity in social networks. Proceedings of the Second Edition of the ACM Conference on Online Social Networks - COSN '14, 83–94. https://doi.org/10.1145/2660460.2660467

Peddinti, S. T., Ross, K. W., & Cappos, J. (2017a). Mining Anonymity: Identifying Sensitive Accounts on Twitter. ArXiv:1702.00164 [Cs]. Retrieved from http://arxiv.org/abs/1702.00164

Peddinti, S. T., Ross, K. W., & Cappos, J. (2017b). User Anonymity on Twitter. IEEE Security & Privacy, 15(3), 84–87. https://doi.org/10.1109/MSP.2017.74

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2). https://doi.org/10.18637/jss.v048.i02

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. The Journal of Educational Research, 99(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Suler, J. (2004). The Online Disinhibition Effect. CyberPsychology & Behavior, 7(3), 321–326. https://doi.org/10.1089/1094931041291295

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). Using multivariate statistics (Seventh edition). Boston: Pearson.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. Australasian Journal of Paramedicine, 8(3). https://doi.org/10.33151/ajp.8.3.93

Xue, M., Yang, L., Ross, K. W., & Qian, H. (2017). Characterizing user behaviors in location-based find-and-flirt services: Anonymity and demographics: A WeChat Case Study. Peer-to-Peer Networking and Applications, 10(2), 357–367. https://doi.org/10.1007/s12083-016-0444-5

## Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 740773.

This research has been funded by the Ministry of Science, Innovation and Universities under FPU Grant Reference FPU16/01671.

This research has been funded by the Spanish National Cybersecurity Institute (INCIBE) under "Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad" Grant Reference INCIBEI-2015-27349.

# LISTA DE TABLAS

# LISTA DE FIGURAS