DETECCIÓN AUTOMÁTICA DE LA MISOGINIA EN TWITTER

TRABAJO FIN DE MASTER

Autor

David Rubio Martínez

Tutora

Laura Plaza Morales

Máster en Tecnologías del Lenguaje

Curso Académico 2019-2020

Madrid - Septiembre 2020

DETECCIÓN AUTOMÁTICA DE LA MISOGINIA EN TWITTER

David Rubio Martínez

Máster en Tecnologías del Lenguaje
Curso Académico 2019-2020

Septiembre de 2020

Agradecimientos

En primer lugar, me gustaría agradecer a mis padres su apoyo incondicional para continuar mis estudios de postgrado y por transmitirme la necesidad de tener una formación constante, estando al día de las diferentes tecnologías que se nos presentan con el paso del tiempo. Sin duda, su apoyo y confianza han sido claves para poder afrontar este curso.

Agradecer también a todos los profesores de las distintas asignaturas del Máster su buena disposición para ayudarme cuando ha sido necesario, en especial a Laura Plaza por haberme ofrecido este Trabajo de Fin de Máster y por su apoyo y ayuda durante la realización del mismo.

Adicionalmente, me gustaría agradecer también a mis compañeros de trabajo su apoyo a lo largo de todo el curso y la plena confianza que todos han depositado en mí. Su apoyo diario ha sido uno de los pilares básicos para poder haber terminado de forma exitosa el curso.

Por otro lado, agradecer a mi novia su confianza en mí a lo largo de todo el curso y por tener unas expectativas tan altas sobre la consecución de mis objetivos.

En general, gracias a todos los que nunca habéis dudado de mí, ya que, gracias a vuestras muestras de apoyo, me han permitido no dudar de lo que soy capaz.

"Silent gratitude isn't much use to anyone" - Gertrude Stein

Resumen

En la actualidad, podemos observar cada vez más episodios de acoso en las redes sociales. El crecimiento exponencial de estas tecnologías también ha favorecido el desarrollo de comportamientos inapropiados y malintencionados, donde podemos destacar de una forma genérica el hate speech, término que utilizamos para referirnos a casos de agresividad y discurso ofensivo en contra de un determinado grupo social, ya sea por cuestiones de raza, sexo, religión, orientación sexual etc.

A lo largo de este trabajo analizaremos el hate speech y las consecuencias que se pueden derivar de ese comportamiento. Una vez analizado, haremos especial hincapié en un tipo de hate speech que se produce contra las mujeres, esta es la misoginia online. Desgraciadamente, el odio en contra de las mujeres está teniendo lugar cada vez con más frecuencia, destacando que la misoginia puede expandirse sin control alguno.

Podemos definir la misoginia como el odio o prejuicio hacia la mujer, pudiendo ser manifestada lingüísticamente en diversas formas, incluyendo exclusión social, discriminación, hostilidad, amenazas violentas y objetificación sexual.

La facilidad de difusión, el anonimato y la sensación de carencia de penalizaciones han dado lugar a que las redes sociales se conviertan en plataformas de ataque contra ciertos colectivos con diversos fines.

Una de las plataformas en las que más activamente está teniendo lugar este tipo de comportamientos es Twitter, red social que permite compartir textos de hasta 280 caracteres. En este ámbito, surge la necesidad de contar con herramientas y sistemas que nos permitan detectar de una forma automática comportamientos misóginos en las diferentes redes sociales. Este trabajo de fin de máster tiene por objetivo la detección automática de la misoginia en la plataforma de Twitter.

Gracias a las técnicas de procesamiento de lenguaje natural y a las diferentes tecnologías emergentes que han ido apareciendo en los últimos años, se ha favorecido la detección de este tipo de ataques en contra de la mujer, pero todavía queda mucho trabajo por realizarse y es un área de estudio en el que el estado del arte no se encuentra todavía en una etapa consolidada.

Repasaremos distintos estudios que se han realizado sobre sistemas de detección del hate speech y la misoginia online para posteriormente desarrollar un sistema de detección de la misoginia en Twitter y que al mismo tiempo nos permita realizar una clasificación de la misma en diferentes categorías.

Finalmente, se propondrá cuáles pueden ser las próximas líneas de investigación en materia de detección de la misoginia en las redes sociales, ya que a pesar del creciente número de estudios e investigaciones que se están realizando en esta materia, todavía queda un amplio margen de mejora en el desarrollo de sistemas de detección de la misoginia online.

Índice general

Índice de figuras

Índice de tablas

Acrónimos Utilizados

IT	—	Information Technology
IOT	—	Internet Of Things
AMI	—	Automatic Misogyny Identification
PRC	—	Pew Research Center
ICCPR	—	International Covenant on Civil and Political Rights
FTA	—	Face Threating Acts
CMC	—	Computer Mediated Communication
DUDH	—	Declaración Universal de Derechos Humanos
CERD	—	Committee on the Elimination of Racial Discrimination
CEDAW	—	The Convention on the Elimination of all Forms of Discrimination Against Women
PIDCP	—	Pacto internacional de derechos civiles y políticos
BOW	—	Bag of Words
POS	—	Part Of the Speech
SVM	—	Support Vector Machine
SASM	—	Semantic Analysis on Social Media
DTM	—	Document Term Matrix
LDA	—	Latent Dirichlet Allocation
LSTM	—	Long Short Term Memory
GRU	—	Gated Recurrent Unit
RNN	—	Recurrent Neural Network

1
Introducción

1.1. Motivación

No cabe duda de que la aparición de la Web en 1991 supuso un gran avance en cuanto a la accesibilidad y disponibilidad de la información. No es solo la cantidad de información disponible, sino lo fácilmente accesible que se encuentra. En los últimos años, se ha producido un crecimiento exponencial de la cantidad de información disponible en Internet.

La web ha evolucionado desde su creación de forma rápida en diferentes aspectos:

- Rapidez de acceso y número de usuarios conectados.
- Ámbitos de aplicación. El uso de Internet ha aumentado exponencialmente desde su creación, actualmente, múltiples de las actividades cotidianas que realizamos se pueden realizar de forma más rápida y eficaz a través de la web.
- Tipo de interacción con el usuario. La evolución que ha seguido la web en relación al rol que los usuarios tienen en el acceso a la misma ha ido también evolucionando. Esta evolución viene principalmente motivada por el cambio de la web estática, donde se mostraba siempre la misma información en todo momento, a la web dinámica, enfocada en la experiencia del usuario, ofreciendo interacciones y caracterizándose por estar en constante desarrollo, actualizándose con frecuencia.

Todos somos conscientes de los beneficios que la Web nos ofrece, pero ante tal cantidad de información y fuentes de datos diferentes, debemos de ser capaces de tratar de forma correcta la información que se nos presenta y de extraer el contenido realmente válido y útil.

La proliferación y expansión de la Web ha provocado la aparición de nuevas plataformas, tales como las redes sociales. Una red social es una estructura compuesta por un conjunto de usuarios que están relacionados de acuerdo a algún criterio.

A lo largo de los últimos años, se puede observar una expansión de estas redes sociales en cuanto a su utilización y volumen de usuarios, siendo una herramienta de uso cotidiano. A pesar de todos los beneficios que nos aportan estas nuevas tecnologías, a menudo vemos otras desventajas en ellas, como puede ser la aparición de *bullying*, ataques o incluso amenazas a personas en la red que pertenecen a distintos grupos sociales.

En concreto, observamos más y más episodios de acoso en contra de la mujer en las redes sociales, ya sean dirigidos a este grupo en general o centrado en una persona concreta. Por lo tanto, cada vez se vuelve más necesario identificar casos de agresividad y discurso ofensivo (*hate speech*) en contra de la mujer.

Durante los últimos años, se le ha prestado más atención al rol de la mujer dentro de la sociedad. Desafortunadamente, debido en muchas ocasiones a casos de odio. De acuerdo al informe en 2017 del Centro de investigación de acoso online (Duggan, 2017) podemos destacar que el 41 % de las personas fueron atacadas personalmente, de las cuales 18 %

fueron objeto de serios acosos a causa de su género, y se concluyó que las mujeres sufren más ataques que los hombres (11 % vs 5 %) en Internet.

Podemos definir la misoginia como el odio o prejuicio hacia la mujer, pudiendo ser manifestada lingüísticamente en diversas formas, incluyendo exclusión social, discriminación, hostilidad, amenazas violentas y objetificación sexual (Anzovino et al., 2018)

Debido al crecimiento exponencial de las redes sociales y las plataformas de microblogs, como puede ser Twitter, el odio en contra de las mujeres está teniendo lugar cada vez con más frecuencia, favoreciendo a la expansión de la misoginia sin control alguno.

Más concretamente, la motivación que me lleva a realizar un trabajo de fin de máster relacionado con el tema propuesto se basa en lo siguiente:

- En primer lugar, la necesidad de detectar los comentarios que puedan ser ofensivos para el colectivo femenino, ya sean dirigidos a este colectivo en general o a una persona en concreto. Aún en el siglo XXI, existen comportamientos que no se deben de tolerar y se deben de tomar medidas respecto a ellos.
- La cantidad de información disponible en la Web ha crecido exponencialmente en la última década, por lo que podemos extraer una información muy útil que nos ayude en nuestras investigaciones y estudios, pero, por otro lado, al existir tal cantidad de información, nos supone un reto, ya que debemos de ser capaces de tratar la información utilizando las herramientas correctas, utilizando procedimientos adecuados y contrastando la veracidad y actualización de los datos.
- Twitter es un servicio de microblogs donde todo el mundo puede expresar sus opiniones y es una enorme fuente de información. Existen opiniones de todo tipo, por lo que nos encontramos con multitudes de tweets relacionados con la misoginia.
- Al desarrollar un sistema para el reconocimiento automático de la misoginia, podremos detectar aquellos tweets que menosprecian el papel de la mujer en la sociedad y su motivación, lo cual nos ayudaría a intentar ponerle freno a esta clase de comentarios.
- La puesta en práctica de técnicas que hoy en día reciben un amplio interés y atención por parte de la comunidad científica, como son *machine learning* y el procesamiento de lenguaje natural.
- La utilización de la tecnología como herramienta para solucionar problemas existentes en la sociedad y su contribución a erradicar comportamientos inapropiados.

En resumen, este trabajo de fin de máster combina dos vertientes que me resultan de especial interés, por un lado, la detección de un problema existente en la sociedad hoy en día, que es la misoginia, y por otro lado, el hacerlo utilizando técnicas de aprendizaje automático (*machine learning*), aplicando para ello lo aprendido a lo largo de este curso así cómo durante toda mi carrera académica.

1.2. Definición del problema

Como se ha definido en el apartado anterior, existe una problemática, que es la misoginia online y su proliferación y expansión en las redes sociales

El estudio de este problema fue propuesto en el congreso IBEREVAL 2018 (IBEREVAL, 2018), celebrado en Sevilla. Se trata de un workshop que fomenta el desarrollo de las tecnologías del lenguaje para el estudio de lenguas íberas (español, portugués, catalán, vasco y gallego) mediante la creación de foros de debate acerca de sistemas de procesamiento de lenguaje natural. La tarea se define en IBEREVAL 2018 cómo AMI (*Automatic Misogyny Identification*).

La misoginia no es algo nuevo que haya aparecido a raíz del crecimiento de la Web, sino que lleva existiendo muchos años, pero con la aparición de nuevas tecnologías y su facilidad de uso, se ha favorecido su expansión, crecimiento y divulgación. Al mismo tiempo que la tecnología ha favorecido su expansión, también nos va a permitir su detección y tratamiento.

Las redes sociales han incrementado su popularidad debido a que permiten a las personas expresarse desde perspectivas personales e incluso anónimas sobre cualquier tema y cuando deseemos.Teóricamente, debería de existir una correlación transparente entre un usuario online y la persona offline, es decir, que la identidad de una persona en la red no debería ser desconocida ni permanecer bajo el anonimato, pero en muchos casos, esta premisa no se cumple. Dado que la Web y las redes sociales permiten la libertad de expresión, algunas personas se benefician de la posibilidad de esconderse detrás de un perfil anónimo para poder atacar y ofender a otras personas y colectivos.

Como consecuencia del creciente aumento de casos de ataques online, los gobiernos solicitaron a las diferentes plataformas de redes sociales una exhaustiva monitorización de su contenido para proteger a los usuarios ante ataques online, al igual que para no permitir comportamientos inadecuados en las plataformas.

Entre todos los grupos sociales que están expuestos a ataques online, podemos encontrar el que es objeto de este trabajo, las mujeres. *Pew Research Center* (PRC, 2017) reportó en 2017 que el 21 % de las mujeres entre 18 y 29 años habían sido testigos de ataques sexuales en Internet. Además, el 53 % de las mismas fueron enviadas imágenes explícitas que no habían solicitado. En este contexto, Bailey Poland (WMC, 2020), una conocida escritora y activista de los derechos de las mujeres, ha profundizado en sus estudios en el concepto de "cibersexismo" (Poland, 2016).

El sexismo es un comportamiento abusivo hacia el sexo opuesto, cuyas razones, de acuerdo con la autora, son los prejuicios basados en el género y el poder social usado con intención de ofender y atacar para preservar beneficios personales. Sin embargo, el cibersexismo debe de ser entendido como la afirmación despectiva de la posición masculina hacia la mujer a través de comunicaciones realizadas por ordenador.

Además, Poland defiende que los ataques online provienen de actitudes negativas y

creencias que tienen su origen en el mundo físico. Por lo tanto, la sexualización y objetificación de las mujeres conducidas por estereotipos en la Web y redes sociales son comunes entre industrias del entretenimiento y se ha demostrado que puede tener consecuencias devastadoras en la mujer, como pueden ser los trastornos alimentarios y problemas psicológicos, entre otros.

En uno de los estudios realizados por Poland, se demuestra y analiza cómo es más probable que los hombres sean los primeros que hablen en un grupo mixto de hombres y mujeres, al igual que son los que más frecuentemente interrumpen la conversación. Estas dos acciones conllevan a que sean los hombres quienes principalmente tengan el control de la conversación, favoreciendo que esta pueda girar en torno a los ideales y prejuicios del hombre.

Como complemento a los análisis realizados por Poland, podemos encontrar estudios acerca de los factores potenciales que podrían conducir a ciertas personas a expresar más fácilmente su agresividad en plataformas online que en el mundo real. Estos estudios formalizan lo que podemos llamar *disinhibition effect* o efecto desatado, siendo de especial interés el análisis realizado por John Suler (Suler, 2004), en el que destaca los siguientes factores como consecuencia de este efecto:

- Invisibilidad: la persona que realiza un ataque online no recibe las reacciones no verbales de otras personas, siendo estas generalmente indicadores de desacuerdo sobre expresiones sexistas.
- Carencia de sincronía: se elimina la inmediatez de las interacciones y debido a la falta de continuidad en las contestaciones parece que los ataques son todos menos en serio.
- Falta de autoridad: la sensación de que no existe una autoridad que regule nuestros comentarios y comportamiento en la Web, hace que las personas no tomen acciones prejuiciadas ni con temor a posibles represalias.
- Anonimidad disociativa: capacidad del usuario para separarse su yo online de su yo offline no asumiendo la responsabilidad de lo que ocurre en la Web.
- Imaginación disociativa: el usuario se sumerge en las redes sociales y tiene la sensación de que no está en el mundo real.
- Introyección solipsista: las personas tienden a interpretar los mensajes ambiguos en base a sus propias esperanzas o miedos.

En este apartado hemos visto la problemática a la que nos enfrentamos y cómo otros autores han intentado buscar los motivos por los que se produce la misoginia online y qué factores influyen en ella. Una vez entendido el problema y las posibles causas por las que se produce, continuaremos el trabajo estableciendo unos objetivos que nos permitan ofrecer una solución para reconocer aquellos comportamientos que puedan ser ofensivos para la mujer, es decir, que nos permitan detectar la misoginia online.

1.3. Objetivos

Como primer objetivo de este trabajo, es necesario destacar la concienciación, por parte de todas las personas, del problema al que nos enfrentamos, ya que la misoginia está expandiéndose a un ritmo exponencial en las redes sociales sin control alguno y es necesario que todos pongamos de nuestra parte para denunciar aquellos comportamientos que consideremos ofensivos y no apoyar los movimientos de *hate speech* que se puedan producir en la Web contra diferentes grupos sociales.

Para conocer a fondo el problema, en primer lugar, es necesario llevar a cabo una labor de investigación y de conocimiento del estado del arte de la misoginia y del *hate speech* en las redes sociales. Con esta investigación podremos conocer los estudios que se han realizado hasta la actualidad y las acciones que han motivado que la misoginia y el *hate speech* hayan proliferado tanto de forma online.

Una vez estudiado el problema al que nos estamos enfrentando, se realizará un trabajo que consistirá en la identificación automática de la misoginia, que tendrá por objetivo distinguir los contenidos misóginos de los no misóginos en Twitter para su posterior clasificación. Finalmente, se clasificará el tweet en diferentes categorías que veremos a continuación.

Este proyecto tiene por propósito y alcance el análisis de Tweets en inglés. Para poder realizar el análisis en otros idiomas sería necesario adaptar las librerías y algoritmos utilizados a cada lenguaje, ya que las librerías utilizadas están optimizadas y desarrolladas para interpretar caracteres en inglés para su posterior procesamiento, y si utilizamos el sistema desarrollado con otros lenguajes, los resultados que obtendríamos no serían tan precisos. No obstante, el enfoque utilizado es genérico e independiente del lenguaje.

Este proyecto estará organizado en dos subtareas, tal y como fueron propuestas en IBEREVAL 2018 (IBEREVAL, 2018):

– Subtarea A: Identificación de la misoginia: discriminación de contenidos misóginos de los no misóginos. Vemos un ejemplo de tweet misógino y no misógino:

OBJETIVO	**TWEET**
MISÓGINO	I've yet to come across a nice girl. They all end up being bit**es in the end
NO MISÓGINO	@Real Madrid, you are a bi*ch!

Tabla 1.1: Ejemplos de tweets misóginos y no misóginos.

  Esta primera tarea nos servirá, en primer lugar, para diferenciar entre un tweet misógino y no misógino, y será la base para el desarrollo de la segunda subtarea.

– Subtarea B: comportamiento de la misoginia y clasificación de los objetivos, entendidos éstos como la persona o colectivo contra el que van dirigidos los ataques. Se realizará un reconocimiento de los objetivos, pudiendo ser personas específicas o grupos de mujeres, identificando el tipo de misoginia en contra de ellas.

  En el caso de que un tweet resulte misógino, se realizará una clasificación del mismo en las diferentes categorías:

- Estereotipo y objetificación. Categorización y simplificación de una idea de una mujer. Descripción de apariencia física de la mujer y/o categorización con estándares básicos y simples.
- Dominación. Se plasma la superioridad del hombre sobre la mujer para destacar la desigualdad de género.
- Daño de imagen. Para eludir la responsabilidad del hombre, se trata de justificar el abuso de la mujer. Un intento para perturbar la conversación y redirigirla en una situación más cómoda para el hombre.
- Ataques sexuales y amenazas de violencia. Describir acciones como peticiones de favores sexuales, ataques de naturaleza sexual, intentos de instaurar la superioridad del hombre mediante ataques físicos y amenazas de violencia.
- Desacreditar. Difamar e intentar desacreditar la imagen de la mujer.

Podemos ver ejemplos de tweets para cada una de las categorías anteriores en la siguiente tabla:

CATEGORÍA DE MISOGINIA	TWEET
ESTEREOTIPO OBJETIFICACIÓN	Women are Good only into the kitchen... #makemeasandwitch Big te**ts and no brain...this is what I want as wife!
DOMINACIÓN	Women are inferior to men... so shut up please!
DAÑO DE IMAGEN	@yesallwomen wearing a tiny skirt is "asking for it". Your teasing a (hard working, taxes paying) dog with a bone. That's cruel #YesAllMen
AMENAZAS DE VIOLENCIA ATAQUES SEXUALES	I wanna touch your te**ts... you can't imagine what I can do on your body Stupid bi**ch I'll put you down on the floor and I'all rape you!

Tabla 1.2: Ejemplos de tweets para cada categoría de misoginia.

Adicionalmente se realizará una clasificación de los tweets en función de los objetivos finales de estos:

- Activo (individual): el texto incluye mensajes ofensivos enviados a propósito contra un objetivo específico.
- Pasivo (genérico): se refiere a mensajes que van destinados a un colectivo, no a individuos específicos (por ejemplo, un grupo de mujeres).

A modo de ejemplo:

OBJETIVO	TWEET
ACTIVO	@JulieB stupid crazy psychopthic woman... you should die...
PASIVO	Women: just an inferior breed!!!

Tabla 1.3: Ejemplos de objetivos.

Como hemos visto, éste trabajo fin de máster tiene por objetivo la clasificación de tweets en ingles en diferentes categorías, prestando especialmente atención a aquellos tweets que clasifiquemos como misóginos.

Además de realizar esta clasificación, uno de los objetivos perseguidos en la realización de este trabajo es la concienciación de que el desarrollo de nuevas tecnologías y su disponibilidad están favoreciendo a la divulgación de la misoginia en la Web sin control alguno, por eso, el desarrollo de sistemas de detección automática tiene un papel muy importante a la hora de frenar esta clase de comportamientos y así evitar la propagación masiva de actuaciones inadecuadas. Es muy importante que tratemos de erradicar este problema entre todos.

2
Estado del arte

Siguiendo la estructura definida en los proyectos de investigación, en este apartado abordaremos el estado del arte del *hate speech* y de la detección de la misoginia online.

El estado del arte es una modalidad de la investigación que permite el estudio del conocimiento e investigaciones acumuladas dentro de un área específica. En general, el estado del arte hace referencia al estado último de la materia en términos de investigación.

Para conocer hasta dónde se ha avanzado en las investigaciones relacionadas con la materia y los diferentes enfoques que se han utilizado, analizar el estado del arte de las áreas que tenemos por objeto su investigación nos puede ser muy útil para partir de una base sobre la que comenzar a desarrollar nuestro trabajo.

En el desarrollo de este proyecto, vamos a analizar dos estados del arte que posteriormente nos servirán para el desarrollo de nuestro sistema de detección de la misoginia en Twitter.

2.1. Hate speech en Internet

El *hate speech* en Internet es un tipo de discurso que tiene lugar online, es decir, en la Web, generalmente en redes sociales, con el propósito de atacar a una persona o grupo de personas basándose en atributos como la raza, religión, origen étnico, orientación sexual, discapacidad o género.

El *hate speech* online se sitúa en medio de múltiples tensiones y controversias, ya que es una expresión de conflictos entre diferentes grupos. Es un ejemplo de cómo las tecnologías poseen un alto potencial de transformación, aportando oportunidades y a la vez retos. Implica un complejo balance entre derechos fundamentales y principios, por un lado, y por otro lado la libertad de expresión.

Como podemos apreciar, *hate speech* es un término muy genérico y controvertido. Numerosos tratados como el Pacto Internacional en Derechos Civiles y Políticos (ICCPR) (Wikipedia, 2020a) han intentado delimitar el perímetro de este término. En el pasado se iniciaron planes de actuación para proporcionar una mayor claridad a este tema, sugiriendo mecanismos para identificar los mensajes de odio, en este contexto podemos destacar el Plan de acción de Rabat (ACNUDH, 2020).

Los principales sitios de Internet como son Facebook, Twitter y Google, han establecido sus propias definiciones de *hate speech*, obligando a los usuarios a cumplir una serie de normas, permitiendo a las compañías limitar ciertas formas de expresión.

La velocidad con la que crece Internet y su alcance y difusión convierte en una tarea muy compleja para los gobiernos establecer normativas en el mundo virtual. Como he citado anteriormente, el *hate speech* se encuentra en un nexo complejo con la libertad de expresión, derechos individuales, grupales y minoritarios, así como con los conceptos de

dignidad, libertad e igualdad.

Uno de los primeros estudios realizados en materia de hate speech fue la teoría de la educación diseñada por (Brown and Levinson, 1987). Este análisis se centra en cómo las interacciones sociales influencian la "cara" de los interlocutores, donde la cara es la imagen pública de cada individuo. Se definen dos tipos de caras, llamadas positivas y negativas. La primera es el deseo de una persona de ser socialmente aceptada, mientras que la otra se trata de preservar la libertad individual ante imposiciones externas. La teoría conceptualiza estrategias que mantienen un equilibro social para hacer cooperar a las personas.

A partir de esta teoría, Culpeper formalizó la teoría de la falta de educación (Culpeper, 1996), donde se definen un conjunto de estrategias cuya finalidad es la disrupción social. En este estudio, se defiende que la persona que posea un mayor control del contexto de la comunicación puede caer en acciones de falta de cortesía o educación, pudiendo limitar la libertad de respuesta de otros participantes y amenazar de forma más dañina y agresiva en el caso de que el interlocutor quiera reaccionar. Esto lo hemos visto presente anteriormente en el caso del intento de dominancia de los hombres en ciertas interlocuciones. Estas estrategias de dominancia y de dominación del discurso, se conocen como FTA (*Face-Threating Acts*).

Sin embargo, más recientemente, (Hardaker, 2010) defiende que ninguna de las definiciones de falta de educación que se habían realizado hasta el momento capturaban exhaustivamente el fenómeno de *trolling* o ataque en las redes sociales. El autor realizó una investigación acerca de la falta de educación y CMC (*Computer Mediated Communication*). En cuanto al análisis de la carencia de respeto, Hardaker propone una serie de definiciones para su aceptación:

1. Parodia o burla: la forma puede aparecer de forma maleducada, pero su objetivo es reforzar la cohesión grupal.
2. Mala educación no maliciosa: una forma de expresión que no lleva intención negativa.
3. Fallo de falta de educación: carencia no planeada de modales educados.
4. Fallo de falta de educación maliciosa: la falta de educación intencional no percibida por el destinario.
5. Falta de educación maliciosa frustrada: el destinatario de la falta de educación contrataca al acto.
6. Grosería instrumental: la ofensa intencional y su transmisión efectiva

Adicionalmente a estas definiciones, Hardaker realizó una definición de *trolling* que hoy en día es universalmente aceptada:

"Un troleador es un usuario CMC que construye la identidad de sinceridad deseando ser parte de un grupo en cuestión, incluyendo sus intenciones o manifestado pseudo intenciones sinceras, pero cuyas reales intenciones son causar disrupción o crear conflictos para los propósitos de su propia utilidad" - Hardaker.

Finalmente, el autor remarca la ausencia de investigaciones acerca de la falta de educación que tiene lugar online. La necesidad de estudios en esta dirección es debido a la posibilidad de capturar comportamientos agresivos y peligrosos como el *trolling* y diseñar herramientas computacionales efectivas para prevenir y para controlar conductas similares.

La proliferación del *hate speech* en Internet es analizada por el Consejo de Derechos Humanos de la ONU sobre asuntos de las minorías, y esto plantea un nuevo conjunto de desafíos. Tanto las plataformas de redes sociales como las organizaciones creadas para combatir el *hate speech* han reconocido que los mensajes de odio difundidos en línea son cada vez más comunes y han generado una atención sin precedentes para desarrollar respuestas adecuadas.

Hoy en día, una de las ventajas de la Web es que el contenido está disponible para todo el mundo, en un periodo de tiempo que el creador de contenido puede definir, pero a la vez es una desventaja para el *hate speech*, ya que este contenido puede permanecer en línea durante mucho tiempo, en diferentes formatos, en múltiples plataformas que se pueden vincular unas a otras. En palabras de Andre Oboler (CEO del Instituto de Prevención de Odio en línea) "*cuanto más tiempo esté disponible el contenido, más daño puede infligir a las víctimas y empoderar a los autores. Si elimina el contenido en una etapa temprana, puede limitar la exposición. Esto es como limpiar la basura, no evita que la gente tire basura, pero si no se resuelve el problema, se acumula y se exacerba aún más*".

Por otra parte, el anonimato puede presentar un desafío para lidiar con este problema, ya que Internet facilita el discurso anónimo. Algunos gobiernos y plataformas de redes sociales han intentado aplicar políticas de nombres reales. Dichas medidas han sido muy controvertidas al interferir con el derecho a la privacidad y su intersección con la libre expresión. La mayoría de los ataques de *trolling* en línea y *hate speech* provienen de cuentas seudónimas , que no son necesariamente anónimas para todos. A pesar de que existen técnicas para conocer a un usuario de la Web, su uso no siempre será alcanzable por todos los usuarios.

Otra complicación es el alcance transnacional de Internet, que plantea problemas de cooperación entre jurisdicciones en relación con los mecanismos legales para combatir el *hate speech*, es decir cada país tiene su propia legislación. Si bien existen tratados entre muchos países, no podemos destacar que se caractericen por su efectividad. Esto es particularmente evidente en el contexto de los EEUU, que alojan una gran parte de los servidores de Internet y tienen un compromiso constitucional profundamente arraigado con la libertad de expresión, por lo que es muy complicado que un dato alojado en EEUU sea expuesto públicamente.

En el marco de la legalidad, podemos enmarcar el hate speech dentro de los siguientes acuerdos o tratados:

– Principios internacionales: la declaración universal de Derechos Humanos (DUDH, 2020) de 1948, recoge que todos tenemos derecho a igual protección contra cualquier discriminación en violación de los derechos recogidos en la declaración y también se incluye el derecho a la libertad de expresión. Existen otros acuerdos internacionales

en los que se penaliza el *hate speech* cómo son Convención para la Prevención y el Castigo del Crimen de Genocidio (Wikipedia, 2020b) (1951), la Convención Internacional sobre la Eliminación de Todas las Formas de Discriminación Racial (Wikipedia, 2020c), ICERD (1969) y la Convención sobre la eliminación de todas las formas de discriminación contra la mujer (Wikipedia, 2020d), CEDAW (1981).

– Pacto Internacional de derechos civiles y políticos (Wikipedia, 2020e) (PIDCP): podemos considerarlo como el instrumento legal más utilizado en los debates sobre el *hate speech*. El artículo 19 establece el derecho a la libertad de expresión y el 20 limita expresamente la libertad de expresión en casos de "defensa del odio nacional, racial o religioso que constituye una incitación a la discriminación, la hostilidad o la violencia". El Comité de Derechos Humanos, el organismo de las Naciones Unidas creado por el PIDCP para supervisar su implementación, consciente de la tensión, ha tratado de enfatizar que el Artículo 20 es totalmente compatible con el derecho a la libertad de expresión. En el PIDCP, el derecho a la libertad de expresión no es un derecho absoluto, pudiendo estar limitado legítimamente por los estados en circunstancias restringidas cuando afecte a derecho o reputación de otros.

Como hemos visto, existen numerosas legislaciones que defienden los derechos de las personas y actúan contra el *hate speech* en Internet, pero por otro lado también existe la propia protección que intenta establecer los intermediarios Web, es decir las páginas de Internet en las que se produce el *hate speech*. Podemos destacar las siguientes:

– Twitter: en 2017 estableció nuevas pautas y políticas hacia el *hate speech*. Hay una página completa en el Centro de ayuda de Twitter dedicada a describir su Política de *hate speech* (Twitter, 2020), así como sus procedimientos de aplicación. La parte superior de esta página dice: "La libertad de expresión significa poco si las voces se silencian porque la gente tiene miedo de hablar. No toleramos comportamientos que acosen, intimiden o usen el miedo para silenciar la voz de otra persona. Si ve algo en Twitter que viola estas reglas, infórmenos". La definición de *hate speech* va desde "amenazas violentas" y "deseos de daño físico, muerte o enfermedad de individuos o grupos hasta insultos repetidos y / o no consentidos, epítetos, tropos racistas y sexistas u otro contenido que degrada a alguien".

– Youtube: define el *hate speech* dentro del siguiente marco (Google, 2020): "Alentamos la libertad de expresión y tratamos de defender su derecho a expresar puntos de vista impopulares, pero no permitimos el *hate speech*. El hate speech se refiere al contenido que promueve la violencia o tiene el propósito principal de incitar al odio contra individuos o grupos en función de ciertos atributos, tales como: raza u origen étnico, religión, discapacidad, género, edad, condición de veterano, orientación sexual / identidad de género."

– Facebook: prohíben contenido que sea dañino, amenazante o que tenga potencial para provocar odio e incitar a la violencia. En los estándares de su comunidad (Facebook, 2020), se explica que "Facebook elimina el *hate speech*, que incluye contenido que ataca directamente a las personas en función de su raza, etnia, origen nacional, afiliación religiosa, orientación sexual, sexo, identidad de género. discapacidades o enfermedades graves" Las políticas de *hate speech* de Facebook son aplicadas por 7.500 revisores

de contenido. Debido a que esto requiere una toma de decisiones difícil, surge una controversia entre los revisores de contenido sobre la aplicación de las políticas ya que pueden surgir numerosos puntos de vista acerca de que una publicación se considere ofensiva o no.

Como podemos apreciar, las grandes plataformas de redes sociales dedican recursos y esfuerzos a tratar de frenar la difusión del *hate speech* en Internet.

2.2. Detección de hate speech en Internet

En relación al estudio y detección del *hate speech* en Internet se han venido realizando investigaciones en los últimos años que nos ayudarán a establecer el estado del arte de este tema y nos orientarán en nuestro análisis:

- (Banks, 2010). En este estudio se examinan las complejidades de la regulación del *hate speech* en Internet tanto desde el marco legal cómo por la parte tecnológica. Se revisan las limitaciones en cuanto a que en muchas ocasiones las leyes son de ámbito nacional y no internacional. También se pone de manifiesto cómo las innovaciones tecnológicas pueden reducir el daño causado por el *hate speech*.

- (Silva et al., 2016) plantean una medida a larga escala para identificar los principales objetivos del *hate speech* en las redes sociales Whisper y Twitter. Desarrollaron y validaron una metodología para identificar *hate speech* en ambas redes sociales. Los resultados identifican las diferentes formas de hate speech y proporcionan un amplio conocimiento acerca del fenómeno y aproximaciones para su detección y prevención.

- (Djuric et al., 2015). Para la detección del *hate speech* en Internet, proponen utilizar modelos de procesamiento de lenguaje natural basados en representaciones distribuidas de palabras cómo inputs para el algoritmo de clasificación y con ello ir construyendo un sistema cada vez más complejo que actúe como detector del discurso.

- (Chetty and Alathur, 2018) analizan cómo el crecimiento de las herramientas de IT han favorecido al desarrollo del hate speech en el área del terrorismo. Cualquier acto intencional dirigido contra la vida de una persona causando daño es conocido como terrorismo. Los autores defienden que el *hate speech* es un tipo de terrorismo ya que a menudo conlleva un incidente o evento causante de terrorismo. En este trabajo, se realiza una revisión del *hate speech* en el marco del terrorismo en diferentes redes sociales.

- (Mathew et al., 2019) examinan la dinámica de la difusión de las publicaciones realizadas por los usuarios en Gab (Gab.com). Recogen un dataset con 21 millones de publicaciones e investigan su difusión. Los autores observaron que el contenido generado por los usuarios que utilizan el hate speech tiende a difundirse más rápidamente y alcanzan una audiencia más amplia en comparación con las publicaciones generadas sin hate speech. Adicionalmente, se analizan los usuarios que participan más a menudo en el discurso del odio y se descubrió que estos están mas densamente conectados entre ellos que si los comparamos con los usuarios normales.

– (Menini et al., 2019) presentan un sistema para monitorizar el cyberbullying combinando un análisis de la red social, los mensajes y su clasificación. Los autores estructuran el módulo de clasificación en un dataset construido sobre mensajes de Instagram y describen el cyberbullying monitorizando la interfaz del usuario.

– (Yang et al., 2019) exponen que las interacciones entre los usuarios en las redes sociales se producen de diferentes formas y pueden incluir textos, imágenes y videos. En este artículo, los autores analizan el reto de identificar automáticamente el hate speech con tecnologías de *deep learning*. Se presentan diversas aproximaciones basadas en la combinación de diferentes técnicas de aprendizaje automático para integrar texto e imágenes.

En general, los comportamientos de *hate speech* en Internet son bastante complicados de capturar y detectar en el marco socio-lingüístico descriptivo utilizando recursos computacionales. Sin embargo, es necesario comenzar a implementar herramientas computacionales más potentes para lograr una detección automática de estos comportamientos.

Uno de los primeros aspectos que tenemos que tener en cuenta es que los términos del *hate speech* no son homogéneos y como Schmidt y Wiegand (Schmidt and Wiegand, 2017) defendieron en sus análisis, los primeros términos utilizados para identificar el objeto de esta investigación son mensajes abusivos, mensajes hostiles y calumnias. Además, estas expresiones lingüísticas empezaron a reunir términos generales de lenguaje abusivo u ofensivo.

Otra complicación a la que nos enfrentamos son los datasets heterogéneos y los corpus empleados. Las investigaciones de textos de *hate speech* consumen mucho tiempo y recursos, puesto que un comentario malicioso sería encontrado por muchos no estrictamente malicioso. Este pequeño límite entre lo que puede ser ofensivo o no, causa dificultades en construir corpus para analizar y desarrollar nuevas herramientas. El *hate speech* puede ser diferente en función de la tipología de ataques para cada grupo social. Un dataset debe de ser etiquetado correctamente bajo el alcance de una particular forma de investigación de hate speech. En esta etapa de etiquetado encontramos dos riesgos: por un lado, el nivel de discrepancia entre anotadores, y por otro lado, que los textos pueden no ser lo que esperamos y necesitaríamos el marco del autor para describir los datos disponibles.

En conclusión, cualquier aproximación de detección automática de *hate speech* variará mucho en función del propósito del dataset y de la elección de figuras representativas. Como hemos comentado, puede ser que un ataque a un objetivo por causa de un tipo de odio, por ejemplo, de religión, tenga expresiones lingüísticas diferentes a un ataque por odio de género.

Schmidez y Wiegand resumen los principales recursos que se pueden utilizar para intentar desarrollar un sistema de clasificación automático del hate speech:

– Herramientas y mecanismos simples superficiales: las herramientas más simples, pero más ampliamente adoptadas son mecanismos y algoritmos como *bag of words y n-character grams (ngrams)*. De hecho, son bastante predictivas incluso consideradas individualmente. Otros tipos de características superficiales son si una URL o

mención se encuentra en un texto o no, la cuenta y representación de puntuación, longitud del texto o la media de las palabras en un comentario o incluso el número de caracteres alfanuméricos.

- Generalización de palabras: para abordar el problema de la escasez de resultados que podría proporcionar el *bad of words*, se pueden implementar técnicas de clustering, como podría ser el algoritmo de Brown (Brown et al., 1992) o una técnica de *topic modeling* LDA (*Latent Dirichlet Allocation*) (Blei et al., 2002) .

- Análisis de sentimiento: el *hate speech* y el análisis de sentimiento están fuertemente relacionados. El primero se caracteriza por tener un sentimiento negativo, y nos podríamos apoyar en léxicos de análisis de sentimiento para detectar comportamientos de *hate speech* basándonos en tipos de sentimientos negativos, como podría ser la ira o la frustración, por ejemplo.

  Se han algunos numerosos estudios con la utilización de este recurso, pero cabe destacar el de (Dennis et al., 2015), en el que los autores diseñan un clasificador que usa técnicas de análisis de sentimiento para identificar y clasificar el *hate speech* en discursos de la web como foros y blogs.

- Recursos léxicos: este tipo de mecanismo utiliza recursos léxicos sobre términos generalmente aceptados como representativos del *hate speech*, establecidos como tales en la etapa de clasificación (Badjatiya et al., 2017).

- Figuras lingüísticas: etiquetado POS (*Part-Of the Speech*) es un ejemplo de característica lingüística. Se ha utilizado en varios análisis, aunque no se ha obtenido resultados muy prometedores. Los resultados obtenidos en estudios muestran que se obtiene una mejor precisión de acierto con la utilización de recursos diferentes a etiquetadores POS (Waseem, 2016).

- Mecanismos basados en conocimiento: la construcción de un corpus simplemente mirando a las palabras clave nos puede limitar en nuestro análisis y puede conducir a utilizar preferencias. Además, de cara a representar en su totalidad el *hate speech*, es necesario tener en cuenta el contexto de todo el texto. (Dinakar et al., 2012), se aproximaron al *hate speech* de LGBT aplicando la ontología de ConceptNet (Liu and Singh, 2004), donde aparecen estereotipos manualmente recogidos de la red social Formspring.

- Metadatos: los metadatos describen el contexto de un texto, por ejemplo, sobre el usuario, tiempo y geolocalización. De hecho, la información histórica sobre el autor de un comentario puede ser útil, en el sentido de que un usuario puede tender a postear frecuentemente acerca del *hate speech*.

- Análisis de distintos tipos de información. En muchos casos, los comentarios de textos están enriquecidos con datos no estructurados como puede ser imágenes o audios y este tipo de información también debería ser analizada. Sin embargo, los estudios de detección de *hate speech* online principalmente están enfocados a texto.

Aunque hay una gran variedad en el diseño y utilización de soluciones, las técnicas de *machine learning* supervisadas son las más comunes y con las que mejores resultados

se han obtenido hasta el momento, principalmente con SVM (*Support Vector Machines*). Podemos destacar los siguientes trabajos:

- (Malmasi and Zampieri, 2017) examinan diferentes métodos para detectar hate speech en las redes sociales. Los autores establecen reglas léxicas para esta tarea aplicando métodos de aprendizaje supervisado, utilizando un dataset obtenido para este propósito. El sistema utiliza caracteres n-grams y palabras n-grams. La mejor precisión obtenida es del 78 % utilizando un algoritmo SVM.

- (Del Vigna et al., 2017) intentan contener y prevenir la alarmante difusión de las campañas de odio en Facebook que surgen a raíz de la utilización del hate speech. En primer lugar, los autores distinguen diferentes categorías para la clasificación del odio. Se utilizan figuras sintácticas, análisis de sentimiento y léxicos embebidos de palabras. Con todo esto, los autores diseñan dos clasificadores basados en diferentes algoritmos: el primero basado en SVM y el segundo en una red neuronal recurrente llamada Long Short Term Memory (LSTM). Los resultados muestran la efectividad de estos dos clasificadores.

Recientemente, incluso métodos de aprendizaje más profundos (*deep learning*) han comenzado a ser implementados para este propósito:

- (Pitsillis et al., 2018) proponen el diseño de un sistema basado en el desarrollo de clasificadores RNN (*Recurrent Neural Network*) e incorporan características asociadas con la información del usuario, tales como su tendencia al racismo o sexismo. Estos datos alimentan a los clasificadores junto con los vectores de frecuencias de palabras derivados del contenido del texto. Se muestra su efectividad con un dataset de 16.000 tweets.

- (Founta et al., 2018) estudian el complejo problema del hate speech siguiendo un enfoque más holístico, considerando los diversos aspectos del comportamiento abusivo en Twitter. Los autores analizan las propiedades textuales y de los usuarios desde diferentes perspectivas de comportamiento de hate speech. Se propone una arquitectura de Deep learning, utilizando una amplia variedad de metadatos disponibles combinándolos con patrones extraídos automáticamente dentro del texto de los tweets, para detectar múltiples normas de comportamiento abusivas que están altamente interrelacionadas. El enfoque propuesto es probado con múltiples conjuntos de datos que abordan diferentes comportamientos abusivos en Twitter. Los resultados demuestran un alto rendimiento en todos los conjuntos de datos.

En resumen, el reconocimiento automático del *hate speech* es una tarea complicada, desde el principio hasta el final. La definición puede diferir de un trabajo a otro, aunque recientemente el término *hate speech* ha sido aceptado extensamente por su uso legal.

Como hemos podido ver en el estado del arte, el discurso ofensivo es un problema muy común y en constante evolución, casi al mismo ritmo en el que evolucionan las plataformas digitales y la tecnología, donde el anonimato es una ventaja para aquellos creadores del *hate speech*.

Está claro que tiene que ser controlado cuando están ofendiendo a una persona o un determinado grupo, pero en numerosas ocasiones nos enfrentamos con la libertad de los derechos de expresión y las normativas vigentes. En algunos casos existen leyes internacionales, pero en cuanto a las leyes nacionales pueden confrontarse unas con otras.

También los propios intermediarios de Internet, como pueden ser Twitter o Facebook, poseen mecanismos y leyes propias en cuanto al *hate speech*, pero igualmente se crean controversias acerca de si un contenido es considerado ofensivo.

Hemos comprobado que existen diferentes enfoques y técnicas para tratar de detectar el *hate speech* de forma automática online pero todavía queda mucho trabajo por realizar.

A continuación, analizaremos el estado del arte de un tipo concreto de *hate speech*, la misoginia.

2.3. Misoginia en Internet

Una vez desarrollado el estado del arte del *hate speech* en Internet, y habiendo contextualizado la problemática existente no sólo con la misoginia sino, en general, con los ataques generados en Internet, estamos en una situación apropiada para definir el estado del arte de la detección de la misoginia online.

Como hemos visto al inicio del proyecto, podemos definir la misoginia como el odio o prejuicio hacia la mujer, pudiendo ser manifestada lingüísticamente de diversas formas, tales como la exclusión social, discriminación, hostilidad, amenazas violentas y objetificación sexual. Se conoce como misoginia a la actitud y comportamiento de odio, repulsión y aversión por parte de un individuo hacia las mujeres. Etimológicamente, misoginia es de origen griego, compuesta por *miseo* (odio), *gyne* (mujer) y el sufijo -*ia* (que significa acción) (Significados, 2020).

Se conoce como misógino al individuo que practica la misoginia, es decir, que siente antipatía u odio por las mujeres. A lo largo de la historia han existido misóginos conocidos e influyentes cómo Aristóteles, Sigmund Freud, Friedrich Nietzsche y Arthur Schopenhauer, entre otros.

La misoginia es una conducta practicada desde las civilizaciones más antiguas, ya que la mujer es vista en algunas culturas como la causa de la tentación y de la perdición del hombre. Podemos ver algunos ejemplos que están relacionados con la religión y la historia (Wikipedia, 2020h):

- Grecia antigua: en los estudios que se han realizado de esta época se habla de que la raza humana convivía de forma pacífica, pero Zeus, motivado por su enfado con Prometeo por robar el secreto del fuego, envía a la humanidad un "mal para su deleite", dicho mal es Pandora, la primera mujer, que cargaba un recipiente que se le prohibió abrir, sin embargo, finalmente acaba abriendo, desatando al mundo todos los males: parto, enfermedad, vejez y muerte.

- Budismo: muchos estudios coinciden en que el budismo exalta moralmente a sus monjes varones, aunque también se le aporta valor a las madres y esposas de los monjes, pero menor que a los varones.
- Cristianismo: varios estudios coinciden que la misoginia en el cristianismo ha sido consolidada en numerosas ocasiones por los padres de la Iglesia, en algunos casos afirmando que la mujer es la "entrada del diablo". Sin embargo, otros estudios no opinan lo mismo y defienden la figura de la mujer en el cristianismo, apoyándose en el principio de que "el amor se basa en un profundo respeto mutuo como principio rector de todas las decisiones, acciones y planes".
- Islam: en el capítulo cuarto del Corán encontramos un texto clave totalmente misógino en el que se afirma rotundamente que los varones tienen autoridad sobre las mujeres y éstas son propiedad de los hombres.

Como podemos ver, el papel de la mujer en el transcurso de la historia ha estado marcado en numerosas ocasiones por muchas dificultades y tratos de inferioridad que están desapareciendo progresivamente y todavía se sigue luchando por la totalidad de la igualdad en algunos aspectos.

A pesar de que los comportamientos no apropiados en contra de la mujer se están reduciendo y cada vez las normativas son más rigurosas, todavía quedan comportamientos inadecuados, como es el caso de la misoginia online.

Como hemos comentado con el *hate speech*, la tecnología está en constante expansión y es una herramienta accesible por todo el mundo que favorece en muchas ocasiones al anonimato y a la facilidad de divulgación y de expansión de un contenido.

Dentro de la misoginia online, podemos destacar un área sobre la que se han realizado varios estudios, siendo esta el "cibersexismo". Es un problema que también se puede englobar dentro del *hate speech* en redes sociales.

En su libro, Poland se enfrenta a los problemas causados por la presencia del cibersexismo en las redes sociales (Poland, 2016), describiéndolo de forma precisa y centrándose en que el problema que se puede producir de forma online, puede afectar "offline" en la vida personal de las personas. En el mundo físico, la misoginia puede producirse en varios formatos, desde chascarrillos y situaciones incómodas hasta ataques sexuales con finales trágicos.

El ataque online que más frecuentemente se produce es lo que en inglés se conoce como *slurring* (chascarrillos), de hecho, es común llamar a una mujer de forma despectiva cuando se producen desacuerdos de opiniones. Estos chascarrillos pueden ser los primeros pasos que sigue un misógino para tratar de hacer sentir a una mujer incómoda, sin embargo, no solo mediante insultos se consigue este resultado, sino que una mujer está constantemente bajo juicio por su cuerpo, siendo esta una víctima constante de objetificación. Esta objetificación puede ser física (estereotipo) pero también aparecen otros movimientos que no atacan a la apariencia física de la mujer sino que quieren recordarle a la mujer su lugar, como puede ser el uso del hashtag #GetBackinTheKitchen.

Por otro lado, como hemos visto en el apartado anterior, el *trolling*, tal y como define (Hardaker, 2010), es una de las tácticas dominantes del *hate speech* y de la misoginia online. Si la mujer intenta resistir estos ataques, pueden aparecer incluso amenazas físicas y ataques sexuales. Estos dos actos del cibersexismo son las formas más amenazantes que una mujer puede recibir.

Generalmente, la misoginia online rara vez tiene la intención de hacer daño o causar estragos en el mundo físico o real, siendo su principal objetivo preservar la dominancia masculina en espacios online, limitando las interacciones online de la mujer, eliminando su libertad de expresión. No obstante, los ataques y amenazas online pueden poner en riesgo la vida de una persona en el mundo real, ya que las personas misóginas que atacan de forma online, no tienen control alguno sobre las consecuencias que pueden provocar con sus ataques. De hecho, cuando aparecen episodios misóginos online u offline y son objetos de opinión pública, se intentan evitar responsabilidades y las consecuencias pueden ser devastadoras.

En relación con las repercusiones offline que se pueden producir, Poland analiza tres principales repercusiones: profesionales, psicológicas y personales. De hecho, hoy en día, un candidato para un puesto de trabajo es buscado en Internet antes de una entrevista, por lo que, aparte de dañar la figura y reputación pública con comentarios inapropiados sobre una persona, puede perjudicarle en las oportunidades de una carrera laboral.

El cibersexismo puede causar efectos psicológicos como dañar la autoestima, producir ansiedad, traumas e incluso suicidios, como en el caso Iveco (Periódico, 2020), donde una mujer se suicidó al difundirse un video de contenido sexual en las redes sociales.

2.4. Detección de misoginia en Internet

Como consecuencia de la expansión de estos tipos de comportamientos, se hace cada vez más necesario la detección de la misoginia en Internet para intentar erradicarla y detectar aquellos comportamientos inadecuados, sobre todo en las redes sociales, al crearse debates en los que las personas, animadas unos por otros, pueden llegar a expresar sus pensamientos más peligrosos relativos a la misoginia.

Por lo tanto, debido a la magnitud de información disponible hoy en día en Internet, es necesario desarrollar sistemas que detecten de forma automática la misoginia online, ya que constantemente la información se modifica y aparecen nuevos contenidos.

En relación a la detección de la misoginia online, se han realizado algunos estudios. Cabe destacar, por su relación con este trabajo, la tarea AMI (*Automatic Misogyny Identification*), celebrada en el workshop (IBEREVAL, 2018), donde se propusieron algunos trabajos muy interesantes en relación con la detección de la misoginia, pudiendo destacar los siguientes:

– (Canos, 2018). Este estudio realizado en la tarea AMI de IberEval 2018 tiene por objetivo identificar casos de agresividad y *hate speech* en contra de las mujeres. Des-

cribe un sistema basado en un modelo SVM (*Support Vector Machine*), desarrollando y analizando los resultados obtenidos.

- (Pamungkas et al., 2018). Se describe la participación del grupo 14-ExLab@UniTo en la tarea AMI en IberEval 2018. Cabe destacar que este grupo fue uno de los que mejores resultados obtuvo. El sistema propuesto consiste en una arquitectura basada en SVM (*Support Vector machine*) y explora el uso de varios conjuntos de características, incluyendo un rango de términos amplios basados en el uso léxico de palaras abusivas o malsonantes, prestando especial atención a términos sexistas y palabras abusivas atacando a mujeres tanto en inglés como en español.

- (Shushkevich and Cardiff, 2018). El artículo muestra el sistema desarrollado por los autores para la tarea AMI en IberEval 2018. La técnica propuesta está basada en combinar varios clasificadores simples en un modelo más complejo, que clasifica la información teniendo en cuenta las probabilidades de pertenecer a clases calculado por los modelos más simples. Utilizaron cómo clasificadores las regresiones lógicas, el método de Naive Bayes y clasificadores SVM.

- (Frenda et al., 2018). Se presentan los resultados obtenidos en IberEval 2018 para la tarea AMI. Proponen una aproximación basada en características recogidas a través de la información de sentimiento y un conjunto de léxicos construidos mediante la examinación de tweets misóginos de los datos de entrenamiento proporcionados por los organizadores. Considerando el contexto del tweet, se tiene en cuenta el lenguaje informal utilizado en las redes sociales cómo *slangs*, abreviaciones y hashtags en inglés y en español. Una vez realizado este primer análisis, se utiliza un algoritmo SVM (*Support Vector Machine*) y una técnica de conjunto, obteniendo buenos resultados.

Además de los trabajos presentados en IBEREVAL, podemos destacar otros estudios relacionados con la detección de la misoginia online, tales como:

- (Ghanem et al., 2019) examinan la problemática del comportamiento patriarcal y como el hate speech online contra las mujeres tiene grave consecuencias en la vida real. El artículo presenta un enfoque que es capaz de detectar las dos caras del comportamiento patriarcal, la misoginia y el sexismo, analizando tres colecciones de tweets en inglés y obteniendo resultados prometedores.

- (Lynn et al., 2019) estudian el uso de técnicas de *deep learning* para la detección de la misoginia en el Urban Dictionary (Dictionary, 2020), un diccionario en línea de colaboración colectiva para palabras y frases coloquiales. Se compara el rendimiento de dos técnicas de *deep learning*, LSTM y GRU (*Gated Recurrent Unit*), para detectar el comportamiento misógino con el rendimiento de técnicass de aprendizaje automático mas convencionales, como pueden ser regresión logística, Naive Bayes y *Random Forest*. Se concluye que ambas técnicas de Deep learning examinadas tienen mayor precisión en la detección de la misoginia en el Urban Dictionary que las otras técnicas examinadas.

- (García et al., 2020) analizan en su investigación, la aplicación de tecnologías de análisis de sentimiento y computación social para la detección de tweets misóginos. Por otro lado, compilan el MisoCorpus-2020, un corpus equilibrado sobre la misoginia

> en español, y lo clasifican en tres subconjuntos relacionados con (1) violencia hacia mujeres relevantes, (2) mensajes de acoso a mujeres en español de España y español latino y (3) rasgos generales relacionados con la misoginia. La propuesta combina una clasificación basada en incrustaciones de palabras promedio y características lingüísticas para comprender que fenómenos lingüísticos contribuyen principalmente a la identificación de la misoginia. Se evalúa la propuesta con tres clasificadores de aprendizaje automático, logrando la mejor precisión del 85,175 %.

Como hemos podido comprobar, al igual que se está expandiendo cada vez más el *hate speech* relacionado con la misoginia debido al auge de las redes sociales e información disponible en Internet, también se están desarrollando nuevos sistemas para su detección y tratamiento.

No cabe duda de que estamos en un marco en el que analizar tanta información es complicado, pero con los esfuerzos de la comunidad científica y aparición de nuevos algoritmos y métodos, entre todos podemos detectar comportamientos inadecuados e intentar ponerles freno.

En los siguientes apartados, abordaremos el problema desde el punto de vista teórico y práctico y construiremos un sistema que nos permitirá la detección de la misoginia en Twitter.

3
Twitter Corpus

Para la realización de este trabajo, se ha utilizado el *dataset* de tweets en inglés proporcionado en la tarea AMI del congreso IBEREVAL 2018, citado al comienzo de este trabajo, por lo que se han utilizado los mismos *datasets* de training y *test* que utilizaron los participantes del workshop en 2018.

La construcción de un buen corpus es crucial para el correcto desarrollo de un proyecto de minería de datos. En este caso, se tiene la ventaja de que no es necesario construir un corpus nuevo, sino que ya se dispone de él, gracias a los esfuerzos de los organizadores de IBEREVAL.

El corpus utilizado fue construido a partir de tweets con una longitud máxima de 280 caracteres y sin restricciones en la geolocalización. Los tweets fueron escogidos y descargados siguiendo principalmente tres aproximaciones: selección de un conjunto de palabras representativas, tales como términos singulares, expresiones y hashtags; colecciones de tweets en los que se incluyen una mención al usuario de potenciales víctimas; la descarga de tweets en un determinado tiempo publicados por usuarios misóginos elegidos.

El corpus está formado por dos *datasets* independientes, por un lado, uno de *training* y, por otro lado, otro de *test*. Ambos *datasets* están etiquetados en función de si son misóginos o no, en el caso de que sean misóginos, el tipo de misoginia y por último si son considerados activos o pasivos, tal y como hemos visto en la introducción de este trabajo.

Para los datos de *training* contamos con 4000 tweets, distribuidos de la siguiente manera:

TIPO	**TWEETS**	**CLASIFICACIÓN**	**TWEETS**	**OBJETIVO**	**TWEETS**
MISÓGINO	1785	Estereotipo	179	Activo	1058
		Dominación	148		
		Daño de imagen	92		
		Amenaza	352	Pasivo	727
		Desacreditar	1014		
NO MISÓGINO	2215				
TOTAL	**4000**				

Tabla 3.1: Distribución de datos de *training*

Como se puede apreciar a simple vista, tenemos un mayor número de tweets no misóginos. En cuanto a los tweets misóginos destacan aquellos que podemos clasificar dentro de la categoría "desacreditar", es decir, tweets que intentan difamar y desacreditar la imagen de la mujer.

Adicionalmente, se puede apreciar un mayor número de tweets cuyo ataque se dirige contra un objetivo específico, considerados como "activos".

Por otro lado, para los datos de test, se proporcionan 1000 tweets, distribuidos de la siguiente manera:

TIPO	TWEETS	CLASIFICACIÓN	TWEETS	OBJETIVO	TWEETS
MISÓGINO	460	Estereotipo	140	Activo	401
		Dominación	124		
		Daño de imagen	11		
		Amenaza	44	Pasivo	59
		Desacreditar	141		
NO MISÓGINO	540				
TOTAL	**1000**				

Tabla 3.2: Distribución de datos de test

Podemos comprobar que la distribución es similar a la obtenida para los datos de entrenamiento.

Tenemos una distribución del 80 % para datos de *training* y 20 % para datos de *test*, que en proyectos de analítica de datos suele ser la distribución que se sigue.

Para el desarrollo del sistema trabajaremos con los datos de entrenamiento y una vez lo obtengamos, comprobaremos su eficacia comparándolo con los datos de test.

El formato en el que se nos presentan los *datasets* es tsv (valores separados por tabulaciones) y podemos abrirlo con una hoja de cálculo para un primer análisis valorativo.

Para inspeccionarlo más en detalle sí que es interesante abrirlo con otros programas de analítica de datos, tal y como veremos en los siguientes apartados.

En este trabajo tenemos la ventaja que tanto los datos de *training* como de *test* están etiquetados. En el caso de que no tuviéramos estas etiquetas, el proceso de detección de la misoginia en Twitter sería un trabajo mucho más laborioso y costoso, teniendo que realizar una búsqueda y análisis muy extenso, donde se deberían de utilizar igualmente técnicas de aprendizaje automático, pero en este caso, de aprendizaje no supervisado, como puede ser clustering o *k-means*.

Una vez conocidos los datos de los que partimos, estamos en disposición de pasar a la explicación de la metodología utilizada en el desarrollo del sistema.

4
Metodología utilizada

Para el tratamiento de textos formales, las técnicas tradicionales de procesamiento de lenguaje natural constituyen un mecanismo que generalmente es suficiente para tratarlos, pero para el análisis y procesamiento de microblogs y redes sociales hacen falta mecanismos y técnicas adicionales.

A diferencia de los textos formales, los microblogs poseen un lenguaje no formal desde el punto de vista léxico y sintáctico, con un uso extensivo de abreviaturas y *slangs* (registros coloquiales e informales usados en un idioma) e incluyen elementos no reconocidos en el lenguaje como hashtags, emoticonos, etc., por lo que se requiere que los procesadores lingüísticos estén adaptados.

La minería de datos en redes sociales nos puede aportar un gran valor, ya que podemos llegar a ser capaces de captar y analizar la subjetividad y necesidades del usuario. El conjunto de técnicas para llevar a cabo un procesamiento lingüístico en las redes sociales se conoce como SASM *(Semantic Analysis in Social Media)*. Dentro de las múltiples redes sociales existentes hoy en día (Facebook, Linkedin, Twitter, Instagram o Snapchat entre otros) vamos a centrar nuestro trabajo en Twitter.

SASM es el procesamiento semántico de los textos y de los metadatos para construir aplicaciones inteligentes basadas en datos sociales. SASM involucra: análisis de redes sociales, aprendizaje automático, minería de datos, recuperación de información y procesamiento de lenguaje natural.

El análisis de texto aplicado a las redes sociales tiene un alcance muy amplio, permitiendo, entre otras aplicaciones, entender el sentimiento y la emoción, identificar temas clave, palabras y frases, profundizando en cualquier conversación para comprender qué es lo que la impulsa y cómo ha cambiado el contenido de la conversación a lo largo del tiempo o medir el *share of voice*, analizando el texto para comprender que porcentaje de una conversación tiene que ver con un tema en concreto.

Para comenzar con el desarrollo del sistema, contamos con un dataset de *training* y otro de *test*, los cuales fueron utilizados en IBEREVAL 2018. La distribución de los mismos ya ha sido analizada en el apartado anterior. El formato en el que se nos han presentado ambos *datasets* es .tsv (valores separados por tabulaciones).

Durante la descripción de la metodología utilizada nos centraremos en la identificación de los tweets misóginos ya que la aproximación es similar para la clasificación de la misoginia y distinción de objetivo activo o pasivo, como veremos posteriormente.

A continuación, se muestra un diagrama de las distintas fases en las que se divide la metodología utilizada para el desarrollo del sistema, las cuales desarrollaremos en profundidad a lo largo de este capítulo:

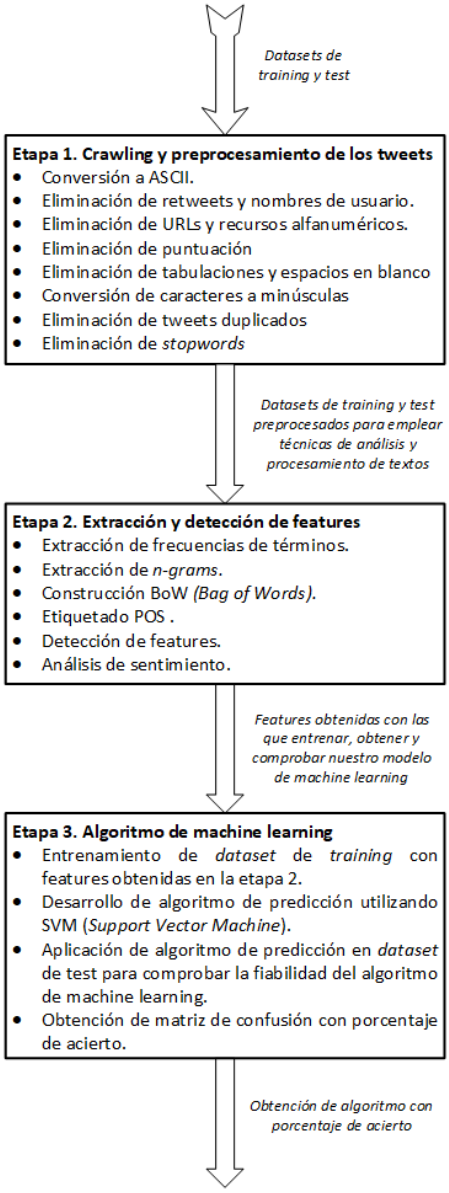

Figura 4.1: Diagrama de la metodología utilizada

4.1. Crawling y preprocesamiento de los tweets

Cuando realizamos un trabajo de procesamiento de datos, prácticamente el 75 % del tiempo dedicado lo emplearemos a tareas de preprocesamiento, donde podemos incluir las actividades de normalización y limpieza de los datos.

Esta etapa es especialmente necesaria en microblogs, como es el caso de Twitter, donde la información se nos presenta de forma no estructurada y con frecuentes errores ortográficos.

En el desarrollo de este sistema, se ha realizado una primera etapa de limpieza y normalización del dataset, que consiste en las siguientes tareas y que es necesario respetar el orden en el que aparecen:

- Conversión a formato ASCII del dataset. Con esta conversión pasamos de tener los tweets presentados en columnas, a disponer de un vector de caracteres para que pueda ser tratado correctamente y evitar problemas de compatibilidad con el software que utilicemos.

- Eliminación de retweets y nombres de usuario. En este trabajo se nos proporcionaron los datasets, y no contienen retweets, pero en el caso de que tuviéramos que descargar nosotros mismos los tweets, la eliminación de tweets repetidos o retweeteados sería algo muy útil para no tener información duplicada. En cuanto a los nombres de usuario, en nuestro estudio no nos van a aportar valor y probablemente los @ de los usuarios nos causen problemas en las etapas de procesamiento del texto. Sin embargo, si nuestro dataset incluyera el autor que ha escrito el tweet, no lo eliminaríamos, ya que podríamos detectar aquellos autores que son más propensos a publicar tweets misóginos.

- Eliminación de URLs. En necesario eliminar las URLs que aparecen en los tweets, ya que no nos aportan información útil y nos causarían problemas a la hora de medir frecuencias de palabras y extraer *features*, dado que nos aparecerían muchas repeticiones de http o https.

- Eliminación de recursos alfanuméricos. Para nuestro estudio no nos interesa tener en cuenta los caracteres numéricos ni tampoco términos en los que se mezclen letras, números y otros caracteres, ya que no aportan valor.

- Eliminación de puntuación. Al igual que los recursos alfanuméricos, los signos de puntuación no nos van a aportar ningún beneficio y pueden ocasionarnos problemas en nuestro procesamiento posterior

- Eliminación de tabulaciones y espacios en blanco. Para no tener más espacios de los necesarios que puedan causarnos problemas a la hora de medir frecuencias, los eliminamos de nuestro vector de caracteres.

- Eliminación de símbolos, ya que este tipo de caracteres no nos van a ser útiles.

- Conversión de todos los caracteres a minúsculas. Este es un paso muy importante, ya que la mayoría de los programas de análisis de datos son sensibles a mayúsculas y minúsculas y es necesario que todos los caracteres estén en el mismo formato, y lo más normal es convertirlos a minúsculas.

- Eliminación de tweets duplicados. Eliminamos información que pueda estar repetida.

- Eliminación de *stopwords*, que consisten en palabras vacías como artículos, pronombres o proposiciones que suelen ser filtradas durante el procesamiento de textos. Los diferentes programas de análisis de datos suelen tener definido por defecto, un conjunto de palabras que forman parte de una variable de *stopwords*.

Con todos estos procesos descritos anteriormente, hemos realizado un preprocesamiento bastante exhaustivo de nuestro dataset.

Como se puede apreciar, la etapa de preprocesamiento de los tweets es un proceso largo que se debe de ejecutar con mucho cuidado, puesto que la precisión de nuestro modelo estadístico dependerá en gran medida de ello.

Una vez que se ha realizado esta etapa, estamos en disposición de buscar aquellos patrones de comportamiento sobre los que construiremos el algoritmo de nuestro sistema de detección de la misoginia.

4.2. Extracción de features

El criterio que utilizaremos para detectar las características de los tweets misóginos será la frecuencia de términos que consideraremos misóginos y que forman parte del *hate speech*, construyendo un *BoW (Bag of Words)* (ML-Mastery, 2019) a partir de palabras y *ngrams* (Wikipedia, 2020f).

Para poder aplicar técnicas de *topic modeling* y de medición de frecuencias, es necesario convertir nuestro actual vector de caracteres en formato ASCII del que disponemos, a un formato corpus, siendo este una lista de listas. En el primer nivel se tienen los documentos, y en cada documento, la primera entrada tiene el texto, y la segunda tiene información del texto de ese documento. Debe de estar compuesto por textos producidos en situaciones reales *(pieces of language)* y la inclusión de los textos que componen el corpus debe de estar guiada por unos criterios lingüísticos explícitos para asegurar que pueda usarse como muestra representativa de una lengua.

El corpus se utiliza principalmente en la etapa de preprocesamiento de los datos, donde los textos se transforman en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis. En esta etapa se define el corpus o conjunto de documentos, siendo representativo y aleatoriamente seleccionados o mediante algún método de muestreo probabilístico evitando duplicidad de documentos.

Una vez que tenemos normalizado nuestro dataset, habiendo eliminado recursos que no nos aportan valor, estamos en disposición de comenzar con la etapa de procesamiento de los datos.

4.2.1. Bag of words y n-grams

Para categorizar y entender el comportamiento de los tweets misóginos, lo haremos basándonos principalmente en dos recursos:

- *Bag of Words (BoW)*: se trata de un recurso utilizado para extraer características de un texto para su uso posterior en modelados, como puede ser para aplicación de

técnicas de *machine learning*. Es una representación de texto que describe la aparición de palabras dentro de un documento. Principalmente, involucra dos partes:

- Un vocabulario de palabras conocidas
- Una medida de presencia de palabras conocidas.

Se llama "bag" of words porque cualquier información acerca del orden o estructura de palabras en el documento es descartada. Este modelo solo se preocupa de si las palabras aparecen en el documento, no dónde.

– *N-grams*: son subsecuencias de n elementos de una secuencia dada. La forma en la que extraemos los gramas se tiene que adaptar al ámbito que estamos estudiando y al objetivo que tenemos en mente. En nuestro caso, hemos construido los *n-grams* sobre la base de las palabras. En concreto, para el desarrollo de este trabajo se han utilizado bigramas (*n-grams* de orden 2), como veremos más adelante.

Antes de comenzar con el análisis de las frecuencias de las palabras y *n-grams*, realizaremos un etiquetado POS *(Part of the Speech)*, para obtener el etiquetado semántico de las palabras y comprobar si existe alguna diferencia en los tweets misóginos de los no misóginos.

Principalmente, el etiquetado lo utilizaremos para determinar la categoría gramatical de cada palabra en una oración (nombre, verbo, adjetivo, adverbio, conjunción, etc.,). La mayoría de los etiquetadores POS utilizan aprendizaje supervisado, incluso algunos etiquetadores amplían el tagset para incluir fenómenos típicos de Twitter, como pueden ser los hashtags, nombres de usuarios o URLs.

El procedimiento para llevar a cabo el etiquetado de los tweets es el siguiente:

– Definición del etiquetado.

– Tokenización (separación de los Tweets en unidades indivisibles).

– Definición del lenguaje utilizado y el tipo de etiquetado.

– Realizamos un *crossword* entre los tokens obtenidos y las categorías gramaticales definidas dentro de la librería utilizada.

Este procedimiento se ha realizado tanto para los tweets misóginos, como no misóginos. En el apartado de resultados obtenidos, veremos las conclusiones resultantes.

A continuación, comenzaríamos a construir nuestro *Bag of Words*, y para ello, es necesario analizar la frecuencia de las palabras, para ello utilizaremos varios métodos.

En primer lugar, dibujaremos, mediante una nube de palabras cuáles son los términos más utilizados en los tweets proporcionados, ya que de forma visual podremos analizar, en una primera aproximación, cuál es el vocabulario utilizado mayoritariamente.

Para obtener el número de veces que se repiten los términos principales, utilizaremos una matriz DTM *(Document Term Matrix)* (ScienceDirect, 2013), la cual es una matriz

matemática que describe la frecuencia de los términos que se producen en una colección de documentos; las filas corresponden a documentos en la colección y las columnas corresponden a los términos. Este tipo de matriz se utiliza principalmente para medir frecuencias en textos.

Haciendo uso de esta matriz, obtenemos, en primer lugar, de una forma analítica y posteriormente con un gráfico, los términos que más se están utilizando en los tweets misóginos.

Además de la utilización de la matriz DTM, cuando realizamos análisis de textos para medir frecuencias de términos, siempre es interesante implementar alguna técnica de *topic modeling*, siendo este un método estadístico para conocer de forma abstracta los temas en los que se pueden clasificar un determinado número de documentos.

En este trabajo se ha utilizado un algoritmo LDA *(Latent Dirichlet Allocation)*, un modelo generativo que permite descubrir tópicos o temas a partir de documentos. Es un algoritmo de machine learning no supervisado que fue presentado inicialmente como un modelo de búsqueda de tópicos por David Blei, Andrew Ng, y Michael I. Jordan (Blei et al., 2003).

LDA considera un documento como una colección de temas, por lo que cada palabra en el documento está considerada como parte de uno o más temas. LDA agrupa o clusteriza las palabras dentro de sus respectivos temas. Como modelo estadístico, LDA proporciona la probabilidad de cada palabra de pertenecer a un tema, y también una probabilidad de cada tema de pertenecer a cada documento.

Para la utilización de LDA, necesitamos escoger un número de temas. Basándonos en el número de temas, LDA separa un documento en palabras. Un número ideal de temas estaría basado en buscar un balance entre granularidad y generalización. Un mayor número de temas puede proporcionar granularidad, pero puede ser complicado dividirlo en temas claramente segregados. Por otra parte, un menor número de temas puede ser generalizado y combinar diferentes palabras de diferentes temas en uno.

En nuestro caso hemos aplicado el algoritmo LDA para diferentes números de temas y los resultados que obtenemos varían en función de dicho número.

Con estas técnicas que se han comentado previamente; nube de palabras *(wordcloud)*, medición de frecuencia de los términos, DTM y *topic modeling* con LDA, hemos obtenido una buena muestra para la construcción de nuestro *Bag-of Words* sobre el que basaremos nuestro algoritmo de detección de la misoginia.

A continuación, realizaríamos el análisis de los *n-grams*. Para ello es necesario seguir los siguientes pasos:

- Asegurar que los tweets que se están analizando tienen el formato char.
- Es necesario crear un vector de caracteres que incluya *stopwords* en inglés. A parte de trabajar con el propio vector que el programa utilizado incluya, es interesante añadir

términos adicionales de forma manual en función del análisis de nuestros datos.

- Se realiza la tokenización de los tweets, dividiéndolos en palabras y eliminamos las *stopwords* que hemos definido previamente.
- En mi caso, he utilizado bigramas ya que creo que es la forma más optima de utilización de *n-grams* para este trabajo.
- Construcción de todos los posibles bigramas en los tweets incluidos en el dataset.
- Una vez que tenemos un dataset que consiste en todos los posibles bigramas, medimos las veces que se repiten, es decir, su frecuencia.
- Obtención de los bigramas que más se repiten en los tweets.

Con los resultados obtenidos, ahora estamos en disposición de construir el recurso *Bag-of Words*, constando de:

- Términos que más se repiten en los tweets.
- Bigramas que más se repiten en el dataset.
- Términos adicionales y hashtags que fomentan la misoginia online. Algunos de ellos han sido obtenidos del trabajo de María Anzovino (Anzovino, 2018), como por ejemplo, *kunt, dyke, pimp, hysterical, #rulesforgirls, #thatswhatslutsdo, #sandwitchmaker*.

Una vez extraídas las *features* léxicas realizando los pasos descritos en este apartado, la próxima tarea será la detección de su presencia en los tweets.

Para ello, es necesario buscar los términos y hashtags de nuestro BoW en los tweets de los que disponemos, tanto en *training* como en *test*. En el caso de que el término aparezca en el tweet, macaremos nuestra variable con un 1 y en el caso de que no, con un 0. Buscaremos todos los términos en todos nuestros tweets.

Como podemos apreciar, nuestros términos de BoW serán nuestras variables y es muy importante aclarar que serán variables booleanas, es decir, solo podrán tener los valores 0 o 1. Esta condición también tendremos que establecer en nuestro *dataset* cuando utilicemos el programa de procesamiento de datos, porque si no, pueden interpretarse como variables numéricas y los resultados que obtengamos no serán fiables.

Para el desarrollo de esta tarea, y siguiendo la definición de BoW, lo que realmente nos interesa es conocer si el término aparece en el tweet o no, sin importar las veces que se repite.

4.2.2. Sentimientos y emociones

Antes de comenzar con el desarrollo de nuestro algoritmo de *machine learning*, se ha realizado una tarea adicional que nos aporta un gran valor añadido, la cual, consiste en la

identificación de los sentimientos a partir de los cuales están escritos los tweets.

Con la utilización del léxico NRC (Emolex, 2016), se ha obtenido una gráfica de sentimiento que reflejan los tweets misóginos. El léxico NRC es una lista de palabras en inglés y sus relaciones con ocho emociones básicas (enfado, miedo, anticipación, confianza, sorpresa, tristeza, disfrute y disgusto) y dos sentimientos (negativo y positivo). Las anotaciones son realizadas de forma manual mediante la colaboración de todo el mundo.

Este léxico es muy utilizado en el procesamiento de lenguaje natural ya que nos ofrece grandes ventajas a la hora de la interpretación de sentimientos.

Gracias a la utilización de este recurso, se ha obtenido una gráfica de sentimiento de los tweets, que se mostrará en el apartado de resultados obtenidos.

4.3. Algoritmo de machine learning

Llegados a este punto, pasaríamos a la parte clave y más interesante de nuestro proyecto, construir nuestro modelo estadístico capaz de detectar la misoginia en Twitter.

Una vez que tenemos nuestro dataset completo y habiendo realizado la detección de *features*, lo separaremos en *training* y *test*. Los primeros datos son los que usaremos para entrenar nuestro modelo estadístico. La calidad de nuestro modelo de aprendizaje automático va a ser directamente proporcional a la calidad de los datos. Por otro lado, los datos de test son los que nos reservamos para comprobar si el modelo generado a partir de los datos de entrenamiento funciona, es decir, si las respuestas predichas por el modelo para un caso totalmente nuevo son acertadas o no.

Normalmente, el conjunto de datos se suele repartir en un 75 % de datos de entrenamiento y un 25 % de datos de test. En nuestro caso, tenemos un 80 % - 20 %.

Para el desarrollo de nuestro modelo predictivo, utilizaremos las técnicas de *machine learning*, en concreto las técnicas de aprendizaje supervisado.

Machine learning o aprendizaje automático es la ciencia cuyo objetivo es desarrollar técnicas que permitan que los sistemas aprendan de forma automática. Es una tecnología que permite hacer automáticas una serie de operaciones con el fin de reducir la necesidad de que intervengan los seres humanos. Lo que se denomina aprendizaje consiste en la capacidad del sistema para identificar una gran serie de patrones complejos determinados por una gran cantidad de parámetros.

Es decir, la máquina no aprende por sí misma, sino un algoritmo de su programación, que se modifica con la constante entrada de datos en la interfaz, y que puede, de ese modo, predecir escenarios futuros o tomar acciones de manera automática según ciertas condiciones. Como estas acciones se realizan de manera autónoma por el sistema, se dice que el aprendizaje es automático, sin intervención humana.

Dentro de *machine learning* debemos de distinguir entre tres tipos principales de aprendizaje automático:

- Aprendizaje supervisado: se basa en la información que disponemos de los datos de entrenamiento. Se entrena al sistema proporcionándole cierta cantidad de datos definiéndolos al detalle con etiquetas. En nuestro caso se nos proporciona la información de si un tweet es misógino o no y su categorización. Una vez que se ha proporcionado la suficiente cantidad de dichos datos, podrán introducirse nuevos datos sin necesidad de etiquetas, en base a patrones distintos que ha venido registrando durante el entrenamiento. Este sistema se conoce como clasificación.
  Otro método de desarrollo del aprendizaje automático consiste en predecir un valor continuo, utilizando parámetros distintos que, combinados en la introducción de nuevos datos, permite predecir un resultado determinado. Este método se conoce como regresión.
  Lo que distingue al aprendizaje supervisado es que se utilizan ejemplos a partir de los que generalizar para nuevos casos.

- Aprendizaje no supervisado: en este tipo de aprendizaje no se usan valores verdaderos o etiquetas. Estos sistemas tienen como finalidad la comprensión y abstracción de patrones de información de manera directa. Este es un modelo de problema que se conoce como clustering. Es un método de entrenamiento más parecido al modo en que los humanos procesan la información.

- Aprendizaje por refuerzo: en la técnica de aprendizaje mediante refuerzo, los sistemas aprenden a partir de la experiencia. Es una técnica basada en a prueba y error, y en el uso de funciones de premio que optimizan el comportamiento del sistema. Es una de las maneras más interesantes de aprendizaje para sistemas de inteligencia artificial, pues no requiere de la introducción de gran cantidad de información.

Para nuestro trabajo de detección de la misoginia, utilizaremos las técnicas de aprendizaje supervisado. Algunas de estas técnicas son *random forest*, redes neuronales o *Support Vector Machines* (SVM). En mi caso, he utilizado SVM para obtener un algoritmo de clasificación y entrenar mi modelo. SVM son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik (Wikipedia, 2020g) y su equipo en los laboratorios AT&T.

Estos métodos están principalmente relacionados con problemas de clasificación y regresión. Como hemos comentado anteriormente, a partir de un conjunto de datos de entrenamiento podemos etiquetar las clases y entrenar un SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, un SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases en 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los dos puntos, de las dos clases, más cercanos al que se llama vector soporte *(support vector)*. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase.

SVM es una técnica de *machine learning* que encuentra la mejor separación posible entre clases. Con dos dimensiones es fácil entender lo que está haciendo. Normalmente,

los problemas de aprendizaje automático tienen muchísimas dimensiones. Así que, en vez de encontrar la línea óptima, el SVM encuentra el hiperplano que maximiza el margen de separación entre clases.

Más formalmente, un SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta, que puede ser utilizado en problemas de regresión. Una buena separación entre las clases permitirá una clasificación correcta.

Podemos ver lo que acabamos de describir con los siguientes ejemplos:

Figura 4.2: *SVM* clasificación

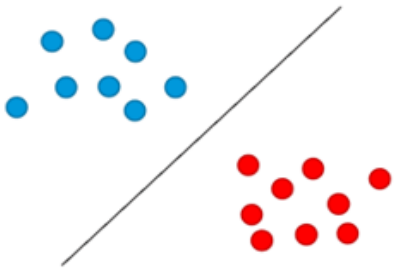

Figura 4.3: *SVM* hiperplano

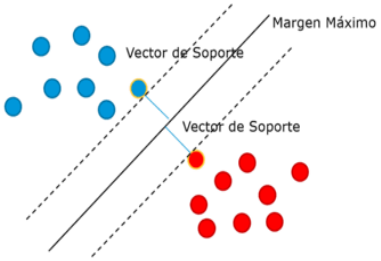

Figura 4.4: Margen y vector de soporte SVM

La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano N-dimensional.

Hay veces en los que no hay forma de encontrar un hiperplano que permita separar dos clases. En estos casos, decimos que las clases no son linealmente separables. Para resolver este problema, podemos usar el recurso o truco del kernel.

Este recurso consiste e inventar una dimensión nueva en la que podamos encontrar un hiperplano para separar las clases. En la siguiente figura vemos cómo al añadir una dimensión nueva, podemos separar fácilmente las dos clases con una superficie de decisión:

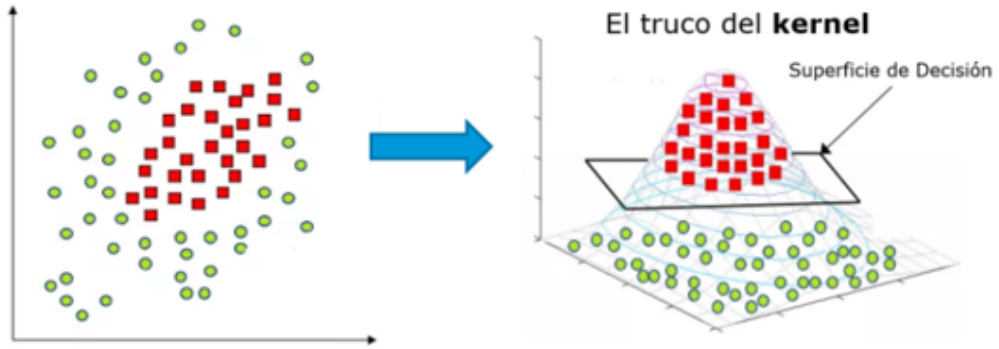

Figura 4.5: Recurso kernel en SVM

Como hemos visto, las técnicas de SVM se basan principalmente en definir un hiperplano, es decir, un subespacio plano y afín de dimensiones p -1. El término afín significa que el subespacio no tiene por qué pasar por el origen. En un espacio de dos dimensiones, el hiperplano es un subespacio de 1 dimensión, es decir, una recta. En un espacio tridimensional, un hiperplano es un subespacio de dos dimensiones, un plano convencional. Para dimensiones p $>$ 3 no es intuitivo visualizar un hiperplano, pero el concepto de subespacio con p - 1 dimensiones se mantiene.

Por lo tanto, la definición matemática para el modelo de SVM, se basa en las ecuaciones de un plano. En el caso de dos dimensiones, el hiperplano se describe acorde a la ecuación de una recta:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0 \tag{4.1}$$

Esta ecuación puede generalizarse para p - dimensiones:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = 0 \tag{4.2}$$

De igual manera, todos los puntos definidos por el vector ($x = x_1, x_2, ..., x_n$) que cumplen la ecuación pertenecen al hiperplano.

Cuando x no satisface la ecuación:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p < 0 \tag{4.3}$$

O bien

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p > 0 \tag{4.4}$$

El punto x cae a un lado o al otro del hiperplano. Así pues, se puede entender que un hiperplano divide un espacio p-dimensional en dos mitades. Para saber en qué lado del hi-

perplano se encuentra un determinado punto x, solo hay que calcular el signo de la ecuación.

Una vez entendido como funciona SVM, estamos en disposición de aplicarlo a nuestro estudio. La función SVM que he utilizado en mi programa de procesamiento de datos, tiene diversos parámetros que se pueden configurar para ajustar nuestro modelo en función de nuestras necesidades.

Para lanzar nuestro primer modelo de SVM que nos sirva como base para conocer un porcentaje de acierto inicial, será necesario indicar en nuestra función:

- Para entrenar nuestro modelo se utiliza el *dataset* de *training*.
- Variable que vamos a predecir y entrenar, en nuestro caso sería *"misogynuous"*, es decir la variable que nos indica si un tweet es misógino (1) o no (0).
- Indicamos a la función SVM que queremos entrenar nuestro modelo con todas las variables de nuestro dataset.
- En tipo de clasificador SVM, elegimos *"C - classification"* dado que nuestra variable a predecir es de tipo factor (0 o 1).
- Utilizaremos el kernel más básico de todos, es decir, el lineal. Con este tipo de kernel, estaríamos separando nuestros datos utilizando una línea, es decir, los tweets misóginos de no misóginos quedan separados linealmente.

La ecuación de este kernel lineal correspondería a la de una recta cuya ecuación sería la misma que hemos visto anteriormente, que también podríamos generalizar de la siguiente manera:

$$k(x, y) = \left(1 + \sum_{j=1}^{p} (x_{ij}, y_{ij})^d\right) \tag{4.5}$$

Donde d sería el grado del polinomio.

Con estos parámetros definidos, obtendremos un algoritmo de *machine learning* con el que entrenaremos nuestro modelo. Una vez obtenido este algoritmo, lo utilizaremos para predecir nuestro *dataset* de *test*, concretamente, la variable "*misogynous*".

Una vez que obtenemos nuestra predicción, utilizaremos una función para obtener la matriz de confusión, que nos permitirá conocer el porcentaje de acierto y ver los falsos positivos que hemos obtenido, es decir, los errores que ha cometido nuestro modelo.

Tal y como y veremos en el apartado de resultados obtenidos, con nuestro modelo obtenemos una precisión bastante alta, pero al ser nuestro primer modelo, tenemos margen de mejora.

La primera mejora para nuestro modelo consistirá en utilizar *cross-validation* con distintos valores del parámetro de tunning C ("*cost*") de la función SVM.

Este parámetro, por defecto tendrá el valor 1 e indica cuánto debe de inclinarse o doblarse el plano o línea en nuestro modelo. Para un valor bajo de *cost*, se obtendrá una línea de baja pendiente y con un valor alto, podremos tener una mayor inclinación.

En la siguiente figura, podemos ver como dependiendo de la inclinación de la línea, la clasificación de los datos varía.

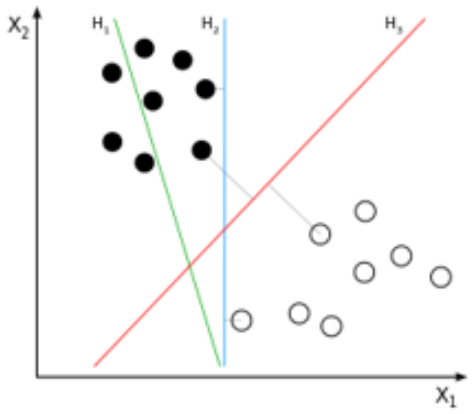

Figura 4.6: Parámetro "cost" SVM

Cuanto más se aproxima C a cero, menos se penalizan los errores y más observaciones pueden estar en el lado incorrecto del margen o incluso hiperplano. C es a fin de cuentas el hiperparámetro encargado de controlar el balance entre *bias* y varianza del modelo.

En cada situación que tengamos que resolver, podremos tener comportamientos diferentes y la inclinación de nuestra línea variará en función de nuestras necesidades, por lo que realizar la tarea de *cross-validation* y probar con diferentes valores de *cost*, nos ayudará a mejorar la precisión de nuestro modelo.

Con este método se mejora la precisión, pero todavía podemos probar a incrementarla utilizando otro tipo de kernel, como puede ser el radial.

El kernel radial es una buena aproximación cuando no es posible separar los datos linealmente. La idea es transformar nuestro plano de datos en un espacio de mayores dimensiones.

Con esta transformación, a nuestra función de SVM, no solo tendremos que pasarle el parámetro C de *tunning*, si no también el parámetro gamma (γ), el cual es un parámetro de ajuste que permite medir la suavidad y los límites de decisión, controlando la varianza del modelo.

Podríamos decir que el valor de gamma (γ) controla el comportamiento del kernel, cuando es muy pequeño, el modelo final es equivalente al obtenido con un kernel lineal, a medida que aumenta su valor, también lo hace la flexibilidad del modelo.

La ecuación del kernel radial para SVM es la siguiente:

$$k(x, y) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - y_{ij})^2\right) \tag{4.6}$$

Si gamma (γ) es muy grande, entonces obtendremos límites de decisión fluctuantes y ondulantes silenciosos, lo que explica la alta variación y sobreajuste.

Si gamma (γ) es pequeño, la línea de decisión o el límite es más suave y tiene poca varianza.

Para verlo de forma gráfica, un kernel radial con un valor óptimo de gamma (γ) tiene la siguiente forma:

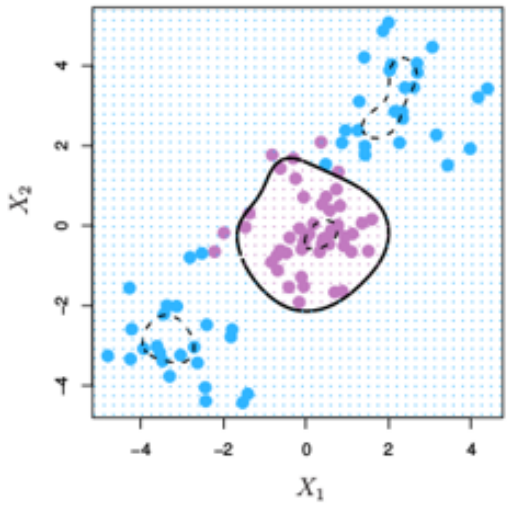

Figura 4.7: SVM kernel radial

Para obtener el algoritmo de SVM lo más preciso posible, se ha utilizado un kernel radial en el que se ha realizado *cross-validation* para buscar el mejor parámetro de C y de gamma (γ), de esta manera se ha mejorado la precisión inicial obtenida mediante la utilización de un kernel lineal con los parámetros por defecto.

Con la obtención de nuestro sistema de *machine learning* basado en SVM, logramos el principal objetivo de este trabajo, que es detectar de forma automática la misoginia en Twitter.

En la descripción de los objetivos de este trabajo, aparecían adicionalmente para aquellos tweets que son considerados misóginos:

– Clasificación de la categoría de misoginia

– Clasificación de objetivo (activo / pasivo).

En el caso de la clasificación de la categoría de misoginia, el planteamiento utilizado ha sido el mismo que el que se ha mostrado para la detección de la misoginia, pero aplicado para cada una de las diferentes categorías: desacreditar, daño de imagen, dominación, estereotipo y ataque sexual.

Para cada tipología, se generará un BoW diferente con el que entrenaremos nuestro modelo basado en SVM.

Para la clasificación del objetivo de los tweets misóginos, la metodología utilizada también es la misma, pero a la hora de construir el BoW con el que entrenaremos nuestro modelo, en el caso de que el objetivo sea activo se ha tenido en cuenta términos que se refieren a una persona en particular y no a un grupo, o lo que es lo mismo, se han utilizado términos y palabras en los que predomina el singular y para los pasivos, el plural.

En los siguientes capítulos veremos los resultados obtenidos durante cada una de las tareas descritas.

5
Herramientas utilizadas

Para el desarrollo de este trabajo se han utilizado los lenguajes de programación R y C#. Los programas en los que se ha desarrollado el sistema han sido principalmente RStudio (https://rstudio.com/) y Visual Studio (https://visualstudio.microsoft.com/es/).

La mayoría de las tareas explicadas en la metodología han sido desarrolladas con la utilización de RStudio. Se ha utilizado Visual Studio con el lenguaje de programación C# para la detección de *features* en los *datasets*.

La elección de RStudio como programa de análisis de datos se debe principalmente a su gran cantidad de funcionalidades y librerías para el procesamiento de textos, así como para la implementación de algoritmos de *machine learning*. La interfaz que nos ofrece R es intuitiva y se divide en varios módulos, por lo que en la misma pantalla podemos ver las variables definidas, el script de R, los resultados obtenidos y las gráficas.

Por otro lado, el motivo de la utilización de Visual Studio se basa en la necesidad de la detección del BoW en los tweets de los *datasets*. Con este software, se ha diseñado un programa en el que introduciremos los tweets en formato .txt que deseemos conocer si son misóginos, y automáticamente obtendremos un fichero de salida .xlsx con variables booleanas, indicando si los términos misóginos extraídos en nuestro análisis, se encuentran en los tweets introducidos o no.

Como podemos apreciar, prácticamente todo el trabajo se ha desarrollado en RStudio. Al inicio del desarrollo de nuestro script no estaremos en disposición de conocer todas las librerías que van a ser necesarias, por lo que conforme avancemos en nuestro desarrollo, será necesario introducir nuevos paquetes.

Las librerías que se han utilizado para el desarrollo de este proyecto se encuentran a disposición del usuario en el repositorio de R del proyecto *The Comprehensive R Archive Network* (CRAN, 2020) y son las siguientes:

Rtweet	*Dplyr*	*Rggobi*	*Pillar*	*RGtk2*
PKgraph	*NLP*	*Tm*	*Methods*	*Tokenizers*
QdapDictionaries	*QdapRegex*	*QdapTools*	*RColorBrewer*	*Qdap*
Textclean	*RJava*	*Rdroolsjars*	*Rdrools*	*RDRPOSTagger*
Worcloud	*Ggplot2*	*Devtools*	*Topicmodels*	*Tidytext*
Textdata	*Quanteda*	*Xlsx*	*Pdftools*	*Stopwords*
Stringi	*Stringr*	*Scales*	*Tidyr*	*Widyr*
GGraph	*Igraph*	*CvTools*	*Lubridate*	*SnowballC*
Tidyverse	*CaTools*	*Caret*	*E1071*	*ElemStatLearn*
Neuralnet				

Tabla 5.1: Librerías utilizadas en RStudio

Como se puede ver, se han utilizado un gran número de librerías a lo largo del proyecto, de las cuales, son de especial importancia:

– *NLP* (CRAN-NLP, 2020). Recoge varios paquetes para la realización de tareas de lingüística computacional en el análisis del discurso y textos en diferentes niveles, enfocándose principalmente en la sintaxis, semántica y pragmática.

- *Tm* (CRAN-TM, 2020). Es un *framework* para aplicaciones de minería de datos en RStudio. Principalmente se utiliza como complemento a la librería NLP.
- *Methods* (RDocumentation-Methods, 2020). Es una librería que proporciona herramientas y funciones para ranking computacional, es decir, se utiliza para conocer la importancia de los términos dentro de un texto.
- *Textclean* (CRAN-Textclean, 2020). Se utiliza para limpieza y preprocesamiento de textos. Su funcionamiento se basa en reemplazar cadenas de caracteres por otras con diferentes variables.
- *RDRPOSTagger* (GitHub-RDR, 2020). Esta librería se utiliza para realizar el etiquetado POS *(Part Of Speech)*.
- *Topicmodels* (CRAN-Topics, 2020). Se utiliza para realizar *topic modeling* mediante LDA *(Latent Dirichlet Allocation)* y CTM *(Correlated Topic Models)*.
- *Quanteda* (CRAN-Quanteda, 2020). Es un framework utilizado para análisis de texto en R. Proporciona funcionalidades para la gestión de un corpus, creando y manipulando tokens y *ngrams*, buscando palabras claves en un texto teniendo en cuenta su contexto y creando diccionarios de contenido.
- *PKgraph* (CRAN-PKGraph, 2020). Es un complemento a las librerías de construcción de gráficas, utilizado principalmente para representar redes neuronales de coapariciones de términos.
- *Caret* (CRAN-Caret, 2020). Esta librería contiene funciones para tareas de *machine learning*, tratamiento de *datasets* de *training* y *test* y modelos para clasificaciones y regresiones lineales.
- *E1071* (CRAN-e1071, 2020). Es una librería utilizada principalmente para análisis de clases. Incluye funciones para clustering, SVM *(Support Vector Machine)* y clasificadores Naive Bayes.

Por otro lado, como se ha comentado anteriormente, para el desarrollo de este trabajo contamos con los *datasets* de *training* y de *test*, pero con la librería *rtweet* podríamos conectarnos con Twitter a través de su API (https://developer.twitter.com/en/apps), donde nos crearíamos una aplicación nueva y podríamos descargarnos tweets que nosotros deseemos que, por ejemplo, incluyeran determinadas palabras o hashtags para analizar si poseen un comportamiento misógino.

Con las claves y contraseñas que se nos presentarían en nuestra aplicación creada en la API de Twitter, estableceríamos la conexión con RStudio y a partir de ahí, podríamos empezar a construir nuestro *dataset* en base a nuestras necesidades.

Al trabajar con RStudio es muy importante que los *datasets* que utilicemos, los carguemos en un formato que R pueda interpretar sin problemas, ya que podríamos tener errores de compatibilidad al implementar las funciones de las distintas librerías. Después de haber cargado todas las librerías de la Tabla 5.1, importaremos nuestros *datasets* de *training* y *test* en formato .xlsx y los guardaremos a continuación en formato .RDS (Report

DataSoruce File), utilizado para almacenar datos en formato nativo a R. Con este paso, evitaremos problemas de compatibilidad de formato a lo largo de nuestro proyecto.

Una vez que se ha comprendido la metodología utilizada y las herramientas con las que se ha implementado, estamos en disposición de conocer los resultados obtenidos durante el desarrollo de nuestro sistema de detección de la misoginia online.

6
Evaluación y resultados

En este capítulo se presenta, por un lado, la metodología de evaluación adoptada para determinar la eficacia de nuestra propuesta de detección de la misoginia online, y por otro lado, los resultados obtenidos, tanto de la evaluación final de la tarea de detección como de las distintas tareas que forman parte del alcance del proyecto, es decir, de la clasificación de la misoginia y su objetivo.

En el primer apartado, se analizará la metodología de evaluación seguida durante el desarrollo de nuestro sistema, donde podremos ver más en detalle el *dataset* utilizado en la tarea. También se plasmarán los resultados que se logran al aplicar la metodología desarrollada en el capítulo 4. De esta manera, se podrán apreciar los diferentes recursos obtenidos, que posteriormente nos han servido para entrenar nuestro modelo estadístico para la detección de la misoginia. En este primer apartado, también se mostrarán las métricas de evaluación de nuestro sistema para todas las tareas relativas a la detección de la misoginia, y podremos comprobar la eficacia de nuestro modelo en cada una de ellas. Como vemos, en esta primera parte, se hace especial hincapié en los resultados obtenidos gracias al procesamiento de los tweets mediante técnicas de procesamiento de lenguaje natural, consiguiendo recursos que posteriormente nos permitirán obtener el principal objetivo de este proyecto, es decir, las métricas de evaluación de nuestro sistema de *machine learning*.

Por otro lado, en el segundo apartado, se mostrarán los resultados obtenidos durante los experimentos realizados a lo largo del trabajo durante las etapas intermedias, tales como *topic modeling* y análisis de sentimiento. Estas tareas no afectan directamente a nuestro algoritmo de predicción de la misoginia pero si que nos ayudan a entender e interpretar mejor el *dataset* al que nos estamos enfrentando y sobre todo, podemos comprobar el tema principal y tópicos que tratan los tweets, al igual que los sentimientos que subyacen en ellos. Finalmente, se interpretan todos los resultados obtenidos a lo largo de las diferentes tareas y su significado, para su correcta interpretación frente al reto presentado inicialmente, es decir, la detección de la misoginia, su clasificación y objetivos a los que se dirige.

6.1. Metodología de evaluación

Como se ha desarrollado a lo largo de este trabajo y plasmado en la figura 4.1, el proceso de evaluación seguido en este proyecto consiste principalmente en la utilización de un algoritmo de *machine learning* basado en SVM, donde utilizaremos nuestro *dataset* de *training* para entrenar el algoritmo de predicción gracias al BoW obtenido, y que lo forman términos misóginos y *n-grams*. Una vez entrenado nuestro algoritmo, comprobaremos su eficacia sobre el *dataset* de test, obteniendo así las métricas de evaluación.

6.1.1. Dataset

Tal y como hemos visto, se han utilizado dos *datasets* durante este trabajo, uno de *training* y otro de test, que fueron proporcionados a los participantes en la tarea AMI (IBEREVAL, 2018), ya comentada a lo largo de este proyecto. Ambos *datasets* están etiquetados y se nos presentan en formato .tsv (valores separados por tabulaciones), que

posteriormente, convertiremos en formato .RDS (Report DataSource File), ya que es un formato nativo a RStudio que nos evitará problemas de compatibilidad en la codificación.

Los *datasets* de *training* y de test se nos presentan separados, pero se podrían unir para la primera tarea de procesamiento de los tweets y separarlos posteriormente de nuevo a la hora de entrenar nuestro modelo con SVM, ya que ambos datasets están completamente etiquetados.

El formato de los *datasets* utilizados a lo largo de este proyecto es el siguiente:

TEXT	**MISOGYNOUS**	**MISOGYNY CATEGORY**	**TARGET**
@emmasharp003 @Ldrake48Lee Bitch shut the fuck up	1	dominance	active
@abzdafab Dear cunt, please shut the fuck up	1	dominance	active
@queenofdragonsb: Pls shut the fuck up bitch	1	dominance	active
@21blvck: "when u gonna get your license"SHUT THE FUCK UP BITCH I AINT GOT TIME DAMN GET OFF MY DICK	1	dominance	active
@SarahhWaqar @CallmeJaagii Bitch shut the fuck up	1	dominance	active
ok bables i'll go to sleep ok bitch shut the fuck up	1	dominance	passive
Your are a fucking ugly bitch! shut the fuck up!!!!! you just a ugly bitch! fuck you!!!!!!!!	1	dominance	active
@bnixole: bitch shut the fuck up you're fucking your best friends dad	1	dominance	active
@vaintshit: shut the fuck up and come suck my dick	1	sexual harassment	active
@AnisaJomha ice has more talent in his ejeculated sperm than you do in ur whole body please shut the fuck up and go suck	1	sexual harassment	active
@ArianasBotch Ok if you fucking said leave block me. But dm me first I'm gonna kick your ass. Shut the fuck up you	1	sexual harassment	active
I heard today from my uncle that if a woman is on her period and she goes into a pool the pool is no better than a toilet. DISGUSTING!	0	0	0
@Collyflowerr no that would be girls WomenSuck	0	0	0

Tabla 6.1: Dataset.

En los *datasets* podemos apreciar si un tweet es misógino (1) o no (0), la categoría de la misoginia y su objetivo. En el caso de que no sea misógino, las columnas "Misogyny

category" y "Target" toman el valor de 0.

Para comenzar con el preprocesamiento de texto y posteriormente su procesamiento, filtraremos por los tweets misóginos inicialmente, ya que nos interesa analizar esta clase de tweets en especial. Una vez que disponemos de un *dataset* en el que solamente tenemos tweets misóginos, obtendríamos los términos que mas se repiten en ellos. A modo de ejemplo, se muestran las palabras con mayor frecuencia:

Término	Frecuencia
bitch	771
women	333
bitch	257
whore	230
cunt	199
hoe	190
ass	186
fuck	179
rape	149

Término	Frecuencia
woman	148
just	146
stupid	135
get	128
fucking	127
dick	120
womensuck	119
can	117
don	117

Tabla 6.2: Frecuencias tweets misóginos

Cabe destacar que, aunque los términos mostrados en la Tabla 6.2 son los más repetidos en los tweets misóginos, para construir nuestro BoW no utilizaremos todos ellos, ya que hay algunos como *get* o *can* que no tendrían cabida en nuestro vocabulario misógino. El filtrado de estos términos se ha realizado manualmente para obtener una mayor precisión y definir un BoW lo más correcto posible y que solo contenga vocabulario misógino que nos asegure una mayor precisión a la hora de obtener un modelo de predicción. Durante esta etapa de filtrado no aparecían muchos términos que no fueran relativos a la misoginia, ni palabras que pudiéramos considerar cómo *stopwords*, ya que en la etapa de preprocesamiento de los datos se han eliminado, y para poder contabilizar los términos se ha realizado un proceso de tokenización, por lo que la eliminación manual de los términos que no nos interesan para construir nuestro BoW, no es una tarea costosa ni lleva asociado un tiempo excesivo de trabajo.

Adicionalmente, en nuestro BoW de palabras misóginas, se han añadido términos que se repiten hasta un mínimo de 50 veces. Esto es debido, principalmente, a que para frecuencias menores de 50, la mayoría de los términos que aparecen no son representativos del ámbito misógino, por lo que a parte de no ser de utilidad para entrenar nuestro modelo de aprendizaje automático, la precisión del mismo se puede ver penalizada.

En el caso de cada categoría misógina, se ha filtrado por un límite de frecuencia mínimo diferente, ya que, para cada caso, la utilización de los términos para construir el BoW varía en función del número de tweets etiquetados en cada categoría, por lo que encontraremos diferentes frecuencias de términos misóginos.

También podemos representar la frecuencia de los distintos términos mediante un gráfico de barras para observar las diferencias en las repeticiones de las palabras.

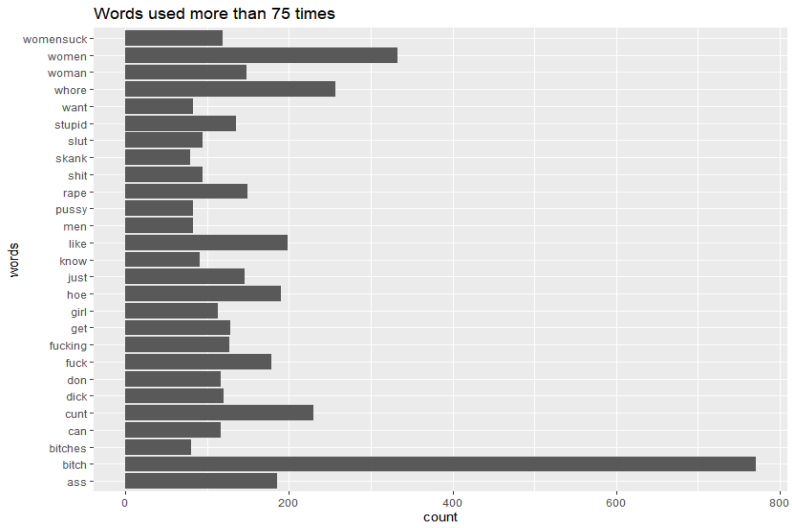

Figura 6.1: Gráfica de frecuencias en tweets misóginos

Otro recurso que se puede obtener para comprobar, en una primera revisión, cuáles son los términos que más aparecen en nuestro *dataset* es una nube de palabras o *wordcloud*:

Figura 6.2: *Wordcloud* misoginia

Si observamos la nube de palabras detenidamente, vemos que la información que se nos presenta, se encuentra acorde con las palabras que hemos visto en la tabla 6.2 que más se repetían en los tweets misóginos: *bitch, women, whore cunt, like, fuck, hoe, womensuck, etc,..*

Podemos apreciar que estos términos misóginos se rigen por las mismas características: vocabulario violento de connotaciones sexuales muy repetitivo, los términos describen la apariencia física de la mujer con estándares básicos y simples, palabras que reflejan la

superioridad del hombre para destacar desigualdad de género y, en general, son términos que se utilizan para intentar desacreditar la imagen de la mujer.

Una vez obtenido los términos que más se repiten a lo largo de nuestro *dataset*, se extraerán los *ngrams* de orden 2 (bigramas) que más veces aparecen en nuestros tweets misóginos, siguiendo los pasos mostrados en el capítulo 4. El resultado es el siguiente:

Bigrama	Frecuencia
stupid bitch	34
ass bitch	33
bitch ass	22
cunt bitch	16
bitch fuck	14
dumb bitch	14
ass hoe	12
look like	12
stupid cunt	12

Bigrama	Frecuencia
dumb cunt	11
fucking cunt	10
stupid ass	10
white women	10
bitch shut	9
whore cunt	9
fucking whore	8
rape culture	8
stupid hoe	8

Tabla 6.3: Bigramas en tweets misóginos

Analizando los bigramas, podemos apreciar que tiene sentido haber obtenido estos resultados, ya que son términos que están profundamente relacionados con el ámbito misógino.

Una vez obtenidos estos *ngrams*, los representaremos mediante un grafo de co-apariciones de términos. Podemos definir un grafo como un conjunto de objetos llamados vértices o nodos unidos por enlaces llamados aristas, que permiten representar relaciones entre elementos de un conjunto. De esta forma, se puede apreciar visualmente cómo se relacionan los diferentes bigramas y palabras que hemos obtenido para construir nuestro BoW.

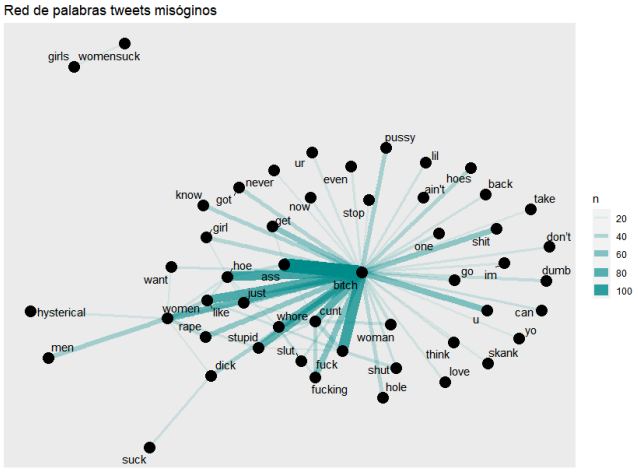

Figura 6.3: Grafo bigramas y términos misóginos

En la figura 6.3 podemos ver como los términos *ass* y *bitch* están unidos por una línea (enlace) de mayor grosor, indicando que los términos (nodos) que están unidos se repiten con mayor frecuencia, y fijándonos en la tabla 6.3, *ass bitch* es uno de los bigramas que más aparece en nuestro *dataset*.

No solo podemos aplicar las técnicas de grafos al análisis de co-apariciones de términos en textos, sino que podemos considerar una red social, por ejemplo Twitter, como una gran red neuronal, en la que se interconectan diferentes personas (nodos) a través de relaciones de amistad (enlaces).

Estos dos ejemplos de redes neuronales en redes sociales nos sirven para entender cómo la teoría de grafos nos puede ayudar a comprender la estructura en la que se basan las redes sociales. Cabe destacar que las propias redes neuronales son un procedimiento de aprendizaje automático supervisado, que también se podría haber utilizado como algoritmo de *machine learning* para la detección de la misoginia online.

Una vez obtenidos los términos y bigramas misóginos que más se repiten en los tweets, y añadiendo algunos términos del ya mencionado trabajo de María Anvozino (Anzovino, 2018), construimos nuestro *Bag of Words*, reflejado en la siguiente tabla:

CATEGORÍA	**PALABRA/S**
TÉRMINOS	*Bitch, whore, cunt, hoe, ass, fuck, rape, stupid, fucking, dick, womensuck, shit, slut, pussy, bitches, skank, hole, girls, hoes, hysterical, shut, dumb*
BIGRAMAS	*Stupid bitch, ass bitch, bitch ass, cunt bitch, bitch fuck, dumb bitch, ass hoe, look like, stupid cunt, dumb cunt, fucking cunt, fucking whore, rape culture, stupid hoe*
TÉRMINOS ADICIONALES	*Slag, fugy, pig, unfuckable, fuckstrugle, lesbian, dyke, grndgrape, pimp, ram, ramming, not all men, over emotional, unreasonable, stay in the kitchen, sandwitch maker.*
HASHTAGS	*#lieastoldbyfemales #ihatefemaleswho #rulesforgirs #mygirldfriendnotallowedto #thatswhatslutdsdo #itsnotrapeif #getbackinthekitchen #mencallmethings #makemeasandwitch #sadwitchmaker #feminismiscancer #notallmen*

Tabla 6.4: *Bag-of Words*

Los términos que se recogen en la tabla 6.4 son las variables con las que entrenaremos nuestro modelo de *machine learning* para la la detección de la misoginia.

Para la realización de este trabajo, hubiera sido interesante haber dispuesto de los usuarios que han escrito los tweets, ya que a parte de construir un BoW con los términos misóginos, hubiéramos añadido una variable más a nuestro sistema, que hubiera sido la tendencia de un usuario de Twitter a publicar tweets misóginos.

Por otro lado, se ha analizado si mediante un etiquetado PoS (*Part of the Speech*) podríamos apreciar diferencias entre los tweets misóginos de los no misóginos, para que nos permitieran hacer uso de ellas, y así entrenar nuestro modelo de aprendizaje automático.

Una vez obtenido el etiquetado de cada una de las palabras que forman los tweets tal y como se ha desarrollado en el capítulo 4, se ha contado el número de palabras de cada categoría gramatical. Es importante que este etiquetado lo realicemos una vez que tenemos normalizado nuestro dataset, ya que si no lo hacemos, es probable que aparezcan composiciones de caracteres sin sentido.

Realizado el recuento, obtenemos lo siguiente:

– Tweets misóginos:

Categoría gramatical	Frecuencia
Adjetivo	2135
Preposición	2196
Adverbio	1721
Auxiliar	1673
Conjunción	712
Determinante	2431
Interjección	144
Nombre	8050
Numérico	1453
Artículo	950
Pronombre	5025
Nombre pronombre	1228
Conjunción subordinada	663
Verbo	4319
Otro	1034

Tabla 6.5: Etiquetado POS tweets misóginos

– Tweets no misóginos:

Categoría gramatical	Frecuencia
Adjetivo	2307
Preposición	2836
Adverbio	2202
Auxiliar	2084
Conjunción	953
Determinante	2825
Interjección	166
Nombre	9064
Numérico	2044
Artículo	1498
Pronombre	5648
Nombre pronombre	1238
Conjunción subordinada	678
Verbo	4948
Otro	2156

Tabla 6.6: Etiquetado POS tweets no misóginos

Si nos fijamos en ambas tablas vemos que no apreciamos diferencias significativas que nos puedan influir en el desarrollo de nuestro sistema, por lo que descartamos el etiquetado

POS como un parámetro con el que entrenar nuestro modelo.

Como hemos podido apreciar, se han presentado los diferentes recursos y estadísticas para la tarea de detección de la misoginia. Para el desarrollo de estas métricas aplicadas a las tareas de clasificación de la misoginia y su objetivo, se han utilizado los mismos pasos y funciones.

Es importante tener en cuenta que para obtener estos recursos para cada categoría de misoginia u objetivo, será necesario filtrar nuestro dataset en función de nuestras necesidades. Por ejemplo, en el caso de que estemos analizando los tweets misóginos de ataque sexual, tendremos que filtrar en primer lugar los tweets misóginos, y posteriormente en categoría de misoginia, seleccionaremos los que pertenecen a ataque sexual. De esta manera, obtenemos para cada categoría de la misoginia las siguientes estadísticas:

Figura 6.4: *Wordcloud* desacreditar

Figura 6.5: *Wordcloud* daño de imagen

Figura 6.6: *Wordcloud* dominación

Figura 6.7: *Wordcloud* estereotipo

Figura 6.8: *Wordcloud* ataque sexual

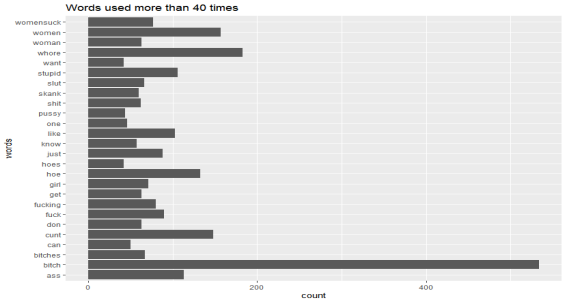

Figura 6.9: Frecuencias desacreditar

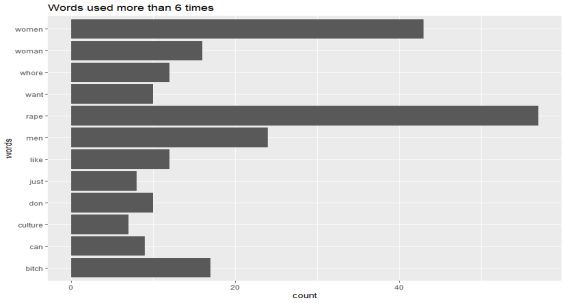

Figura 6.10: Frecuencias daño de imagen

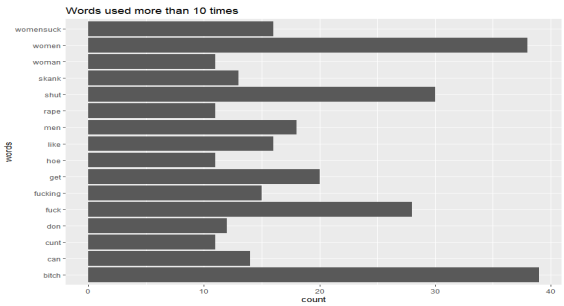

Figura 6.11: Frecuencias dominación

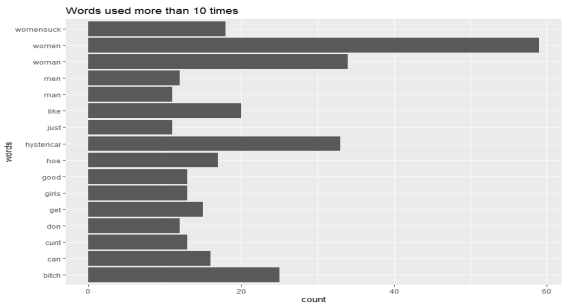

Figura 6.12: Frecuencias estereotipo

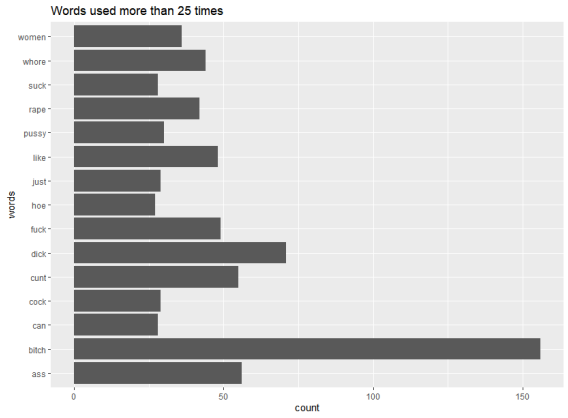

Figura 6.13: Frecuencias de ataque sexual

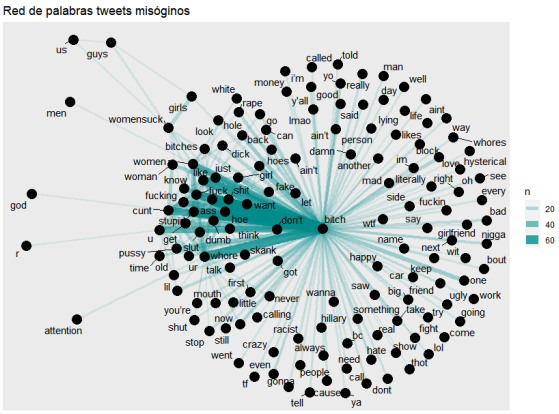

Figura 6.14: Grafo desacreditar

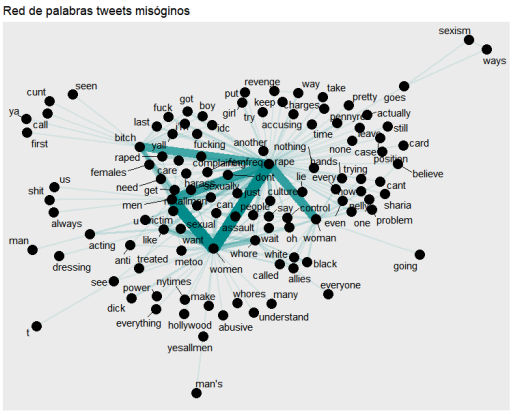

Figura 6.15: Grafo daño de imagen

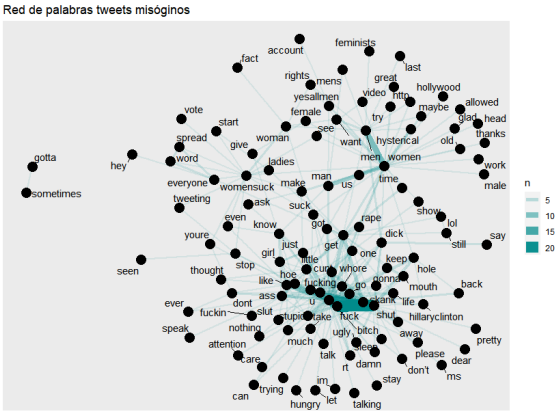

Figura 6.16: Grafo dominación

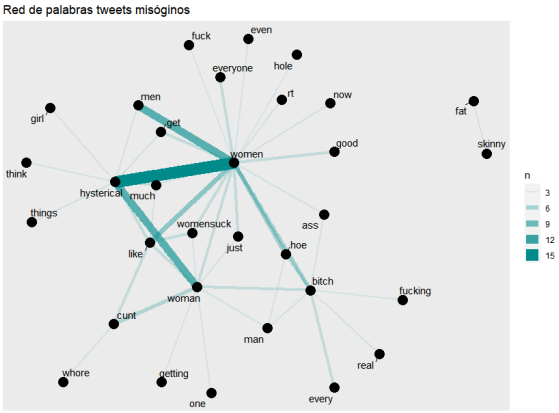

Figura 6.17: Grafo estereotipo

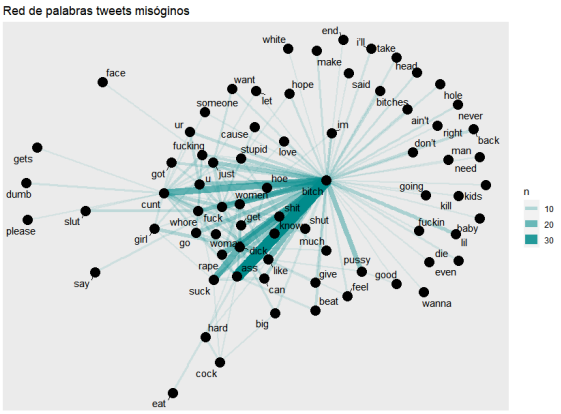

Figura 6.18: Grafo ataque sexual

Tal y como podemos apreciar en las figuras anteriores, dependiendo de la categoría de la misoginia, se obtienen resultados diferentes. Los términos misóginos son parecidos, pero dependiendo de la categoría, se repiten más veces unos términos u otros. A continuación, vemos las diferencias concretas para cada categoría:

- Desacreditar. Los términos que cuentan con una mayor frecuencia son: *bitch, whore, cunt, hoe, stupid, women, womensuck, girl, fuck.*
- Daño de imagen. En esta categoría, vemos que las palabras que más se repiten son: *like, women, men, bitch, culture.*
- Dominación. Podemos destacar el uso de los términos *bitch, fuck, shut, skank, women, hoe.*
- Estereotipo. Para esta categoría, los términos con mayores frecuencias son: *bitch, hysterical, woman, women, womensuck, cunt.*
- Ataque sexual. Las palabras más repetidas son: *bitch, ass, dick, cunt, pussy, rape, whore*

Se puede apreciar que hay palabras que se repiten con mayor frecuencia independientemente de la categoría misógina, como pueden ser *bitch* y *women*, pero hay otros términos que vemos que se repiten un mayor número de veces de forma diferenciadora en determinadas categorías de la misoginia. Por ejemplo, *stupid* en desacreditar, *culture* en daño de imagen, *skank* en dominación, *hysterical* en estereotipo y *rape* en ataque sexual.

Con respecto al grafo de co-apariciones, vemos diferencias entre los distintos grafos que se forman para cada categoría de misoginia, obteniendo también diferentes bigramas.

Las diferencias en cuanto a las frecuencias de términos y bigramas misóginos que se han obtenido para cada categoría, serán clave para construir un BoW distinto para cada una de estas categorías, y así entrenar nuestro modelo de *machine learning* para predecir si un tweet pertenece a un determinado tipo de misoginia.

Igual que se han obtenido los resultados mostrados para las diferentes categorías misóginas, realizamos el mismo procedimiento para los tweets activos y pasivos, es decir, dependiendo de su objetivo. El resultado ha sido el siguiente:

Figura 6.19: *Wordcloud* activo

Figura 6.20: *Wordcloud* pasivo

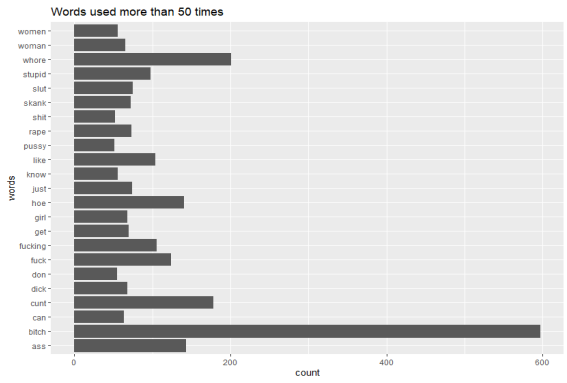

Figura 6.21: Frecuencias activo

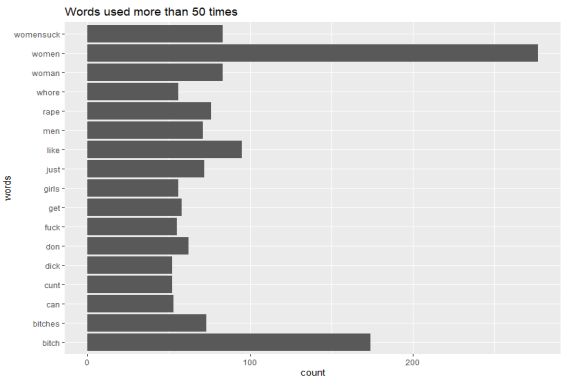

Figura 6.22: Frecuencias pasivo

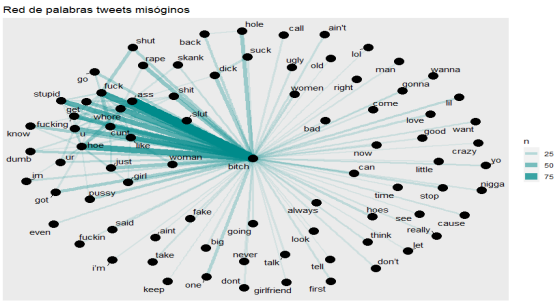

Figura 6.23: Grafo activo

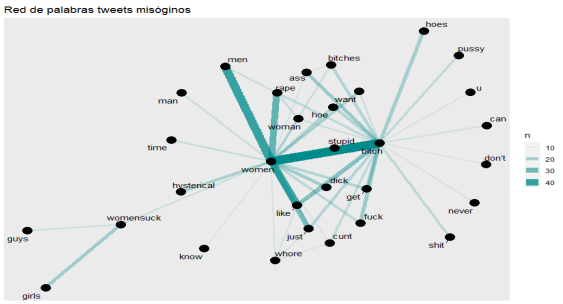

Figura 6.24: Grafo pasivo

De forma análoga a la clasificación de la misoginia, vemos que para los diferentes objetivos, activo y pasivo, se aprecian diferencias en cuanto a la frecuencia de las palabras, reflejadas en los *wordcloud* y las gráficas de frecuencia. Adicionalmente, también vemos que

los grafos de co-apariciones son diferentes, dando lugar a que tengamos diferentes bigramas dependiendo del objetivo misógino. Analizando más en detalle cada categoría:

- Activo. Los términos que más se repiten son: *woman, bitch, hoe, whore, ass, stupid, fuck, slut, cunt*, y en cuanto a bigramas podemos destacar la aparición repetitiva de *ass bitch, stupid btich, cunt bich, dumb bich, stupid cunt.*
- Pasivo. Para esta categoría, las palabras con una mayor frecuencia son: *women, womensuck, girls, bitch, bitches* y respecto a los bigramas podemos destacar *women get, women want, rape culture, beautiful women, women like, black women.*

Si nos fijamos en la lista de términos y *ngrams* para cada objetivo, podemos ver que la principal diferencia entre activo y pasivo radica en que para tweets que ataquen a una mujer en concreto, predominan los términos en singular, y para los que ataquen al colectivo en general, abundan los términos en plural.

Como se ha comentado anteriormente, estas diferencias son muy importantes, ya que nos permitirán construir un BoW diferente para cada objetivo, pudiendo así obtener un sistema clasificador de aprendizaje automático para comprobar si un tweet misógino ataca a una mujer en concreto (activo) o a un grupo de mujeres (pasivo).

Una vez analizadas las estadísticas y figuras obtenidas para el desarrollo de nuestro sistema, estamos en disposición de mostrar las métricas de evaluación definidas para comprobar la eficacia de nuestro algoritmo.

6.1.2. Métricas de evaluación

En este apartado, se describen las métricas de evaluación definidas en nuestro proyecto. Estas métricas se aplican en cada una de las tareas que hemos desarrollado, ya que, tal y como se ha definido a lo largo del proyecto, tenemos una primera subtarea que consiste en la detección de la misoginia en Twitter y una segunda subtarea que consiste en la clasificación de la misoginia y de su objetivo.

La primera métrica de evaluación que se ha definido es la precisión, calculada mediante la siguiente ecuación:

$$Precisión = \frac{número\ de\ instancias\ predecidas\ correctamente}{número\ total\ de\ instancias} \tag{6.1}$$

La precisión nos va a indicar la calidad de nuestro modelo de *machine learning* en tareas de clasificación.

A continuación, las siguientes métricas de evaluación que se han definido, han sido, la métrica F_1 y *recall*.

Comenzaremos definiendo la métrica *recall*. Ésta métrica se refiere a la exhaustividad, la cual, nos va a a informar sobre la cantidad de datos que el modelo de *machine learning* clasifica correctamente. Para calcular esta medida, es necesario conocer la matriz

de confusión, donde se nos presenta, en nuestro caso, el número de tweets que han sido etiquetados correctamente en su categoría y los que se consideran como falsos positivos y falsos negativos, es decir, los que nuestro sistema etiqueta de forma errónea. Esta condición se aplica tanto para la detección de la misoginia, clasificación de tipología y objetivo.

Podemos definir la ecuación de la métrica *recall* de la siguiente forma:

$$recall = \frac{TP}{TP + FN} \tag{6.2}$$

Donde:

TP es el número de verdaderos positivos, es decir, aquellos tweets de una determinada categoría misógina que nuestro modelo etiqueta en dicha categoría de forma correcta.

FN es el número de falsos negativos, es decir, aquellos tweets de una determinada categoría, que el modelo los etiqueta en otra diferente.

Como podemos apreciar, la matriz de confusión es un elemento muy importante en el desarrollo de este proyecto.

Por otro lado, la métrica F_1 se utiliza para combinar las medidas de precisión y *recall* en un solo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

F_1 se calcula mediante la media armónica entre la precisión y la exhaustividad:

$$F_1 = 2\frac{precisión * recall}{precisión + recall} \tag{6.3}$$

En esta ecuación, el valor F_1, asume que nos importa de igual forma la precisión y la exhaustividad, pero esto no tiene porque ser así en todas las ocasiones.

Una vez obtenidos los valores de F_1 y *recall* para cada una de las categorías de la misoginia y su objetivo, obtendremos la métrica Macro F_1, definida de la siguiente forma:

$$F_1 = \frac{F_1(categoría\ de\ misoginia) + F_1(objetivo)}{2} \tag{6.4}$$

Dónde:

F_1(categoría de misoginia) es la media aritmética de las métricas F_1 de cada categoría misógina (dominación, ataque sexual, estereotipo, desacreditar y daño de imagen).

F_1(objetivo) es la media aritmética de las métricas F_1 de cada objetivo misógino (activo y pasivo).

Gracias a esta métrica Macro F_1, obtendremos la precisión y exhaustividad media obtenidas para nuestro clasificador en función del tipo de misoginia y objetivo calculado de

forma conjunta, para establecer una única métrica de evaluación para estas dos tareas.

Con estas tres métricas evaluaremos nuestros diferentes sistemas diseñados para la detección de la misoginia y sus clasificaciones. Los resultados, al igual que su interpretación se desarrollan en el apartado 6.2.3 de este capítulo.

6.2. Experimentación

En este apartado, se detallan los resultados realizados en etapas intermedias del desarrollo de nuestro sistema, como son las estadísticas de *topic modeling* y análisis de sentimiento. Estos recursos no se han utilizado directamente en el desarrollo de nuestro algoritmo, pero nos aportan un gran valor en nuestro estudio acerca de la misoginia.

Por otro lado, se presentan los resultados finales en la detección de la misoginia en las diferentes tareas definidas a lo largo de este trabajo, donde interpretaremos y analizaremos los resultados obtenidos para un correcto conocimiento del funcionamiento de nuestro sistema.

6.2.1. Topic modeling

Como se ha desarrollado en apartados anteriores, en una etapa intermedia de este trabajo se han aplicado las técnicas de *topic modeling*, en concreto LDA, para obtener, de forma abstracta, los temas en los que se pueden clasificar los tweets de nuestro *dataset*.

Para diferentes números de temas, al aplicar LDA sobre nuestro *dataset* obtenemos los siguientes resultados, donde se muestran, a modo de ejemplo, las 10 palabras más repetidas que pertenecerían a cada tema:

- Para 2 temas:
  - Tema 1: *whore, like, stupid, fucking, can, shit, men, skank, girl, think.*
  - Tema 2: *bitch, women, cunt, hoe, ass, fuck, rape, just, get, woman*
- Para 5 temas:
  - Tema 1: *god, believe, work, deserve, ignorant, filthy, trash, scared, ignorant, media.*
  - Tema 2: *act, pretty, fun, phone, respect, new, act, reason, saying, nice*
  - Tema 3: *never, night, die, girlfriend, saw, human, looking, heart, sad, sluts*
  - Tema 4: *women, cunt, hoe, shit, ass, woman, dick, hole, girl, skank*
  - Tema 5: *like, want, fucking, womensuck, rape, stupid, think, get, love, hysterical.*
- Para 10 temas:
  - Tema 1: *wants, boy, enough, hit, feminism, boyfriend, control, feminist, leave.*
  - Tema 2: *getting, ill, stupid, filthy, nobody, comes, foxnews, least, matter, more.*

- Tema 3: *much, didn't, goes, point, post, sorry, text, can, lips, place.*
- Tema 4: *bitch, women, whore, cunt, like, hoe, ass, fuck, rape, woman*
- Tema 5: *sucking, unless, red, fire, death, deserves, nigga, pathetic, whatever, basic.*
- Tema 6: *guys, lie, start, deal, problem, rights, past, balls, facebook, idc.*
- Tema 7: *nasty, men, account, kunt, looks, wear, life, making, since, cares.*
- Tema 8: *also, about, pregnant, acting, true, ffc, mine, tweets, hole, lets.*
- Tema 9: *cum, whole, wrong, stay, went, wit, chill, country, doesn, god.*
- Tema 10: *leave, right, complete, fat, lane, best, condoms, remember, something, cheated.*

Tal y como hemos visto, el número ideal de temas está basado en buscar un balance entre granularidad y generalización. Un mayor número de temas puede proporcionar granularidad, pero puede ser complicado dividirlo en temas claramente diferenciados. Por otra parte, un menor número de temas puede ser generalizado y combinar diferentes palabras de diferentes temas en uno.

En nuestro caso, un buen balance entre granularidad y generalización correspondería a 5 temas o tópicos, al existir una mayor coherencia entre los temas obtenidos y las palabras que pertenecen a dichos temas. Observando los términos que pertenecen a cada uno de estos temas, podríamos representar los tópicos de la siguiente manera:

– Tema 1. Los términos recogidos en este tópico hacen referencia principalmente al odio hacia la mujer en tono despectivo.

– Tema 2. El tópico común de estos términos es la alabanza de la figura del hombre sobre la mujer.

– Tema 3. Estos términos los podríamos enmarcar dentro del tópico de la noche y de los peligros que conlleva para una mujer.

– Tema 4. Se aprecia de forma muy clara que los términos pertenecen al tópico de insultos hacia la mujer, categorizándola con estándares básicos y simples

– Tema 5. Los términos pertenecientes a este tema se pueden clasificar dentro del tópico de sentimientos y comportamientos de la mujer.

Una vez que hemos encontrado nuestro número ideal de tópicos, se ha realizado un análisis de sentimiento de los tweets, el cual aparece desarrollado en el siguiente apartado.

6.2.2. Análisis de sentimiento

Otra de las experimentaciones realizadas a lo largo de este trabajo, ha sido el análisis de sentimiento de los tweets misóginos de nuestro *dataset* mediante la utilización del léxico NRC (Emolex, 2016), y así poder obtener una gráfica con los diferentes sentimientos y sus pesos a partir de los cuales están escritos los tweets misóginos.

Según se ha visto en el desarrollo del estado del arte, hay estudios que han utilizado el análisis de sentimiento como una variable para la detección de la misoginia online, pero sin obtener muy buenos resultados (Plaza et al., 2020).

Aplicando este recurso sobre los tweets misóginos en su conjunto, obtenemos la siguiente gráfica:

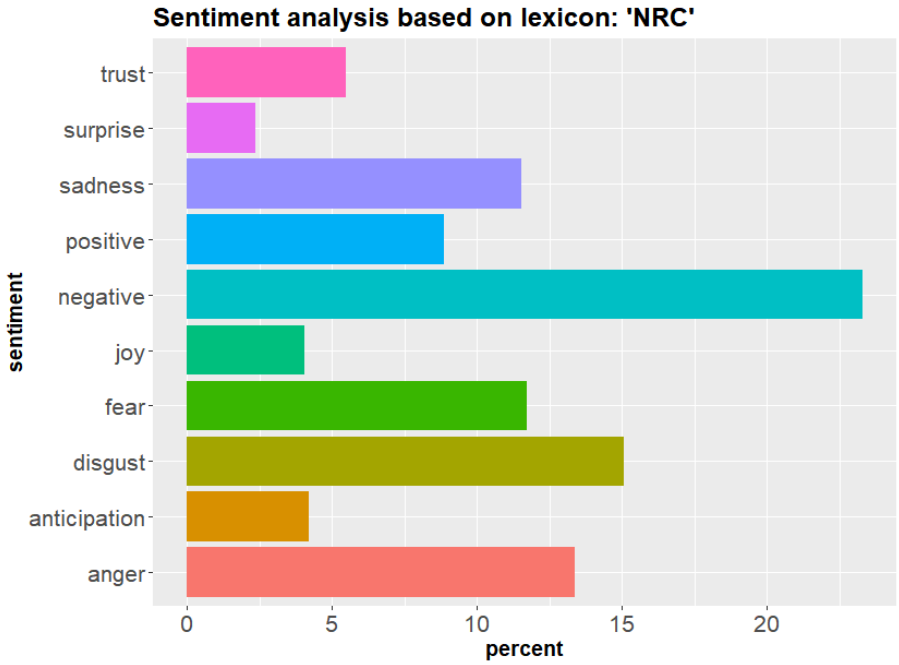

Figura 6.25: Sentimiento tweets misóginos

Como podemos apreciar, la mayoría de los tweets misóginos son escritos partiendo de un estado negativo cuyos sentimientos más comunes son miedo, disgusto, enfado y tristeza, algo normal entre tweets misóginos.

A continuación, se van a plasmar las gráficas de sentimiento obtenidas para las distintas categorías misóginas:

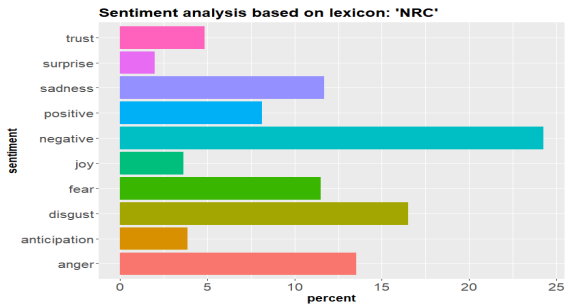

Figura 6.26: Sentimiento desacreditar

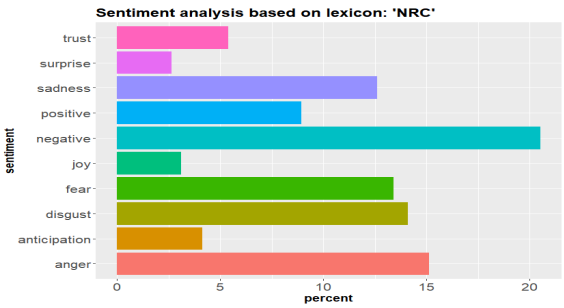

Figura 6.27: Sentimiento daño de imagen

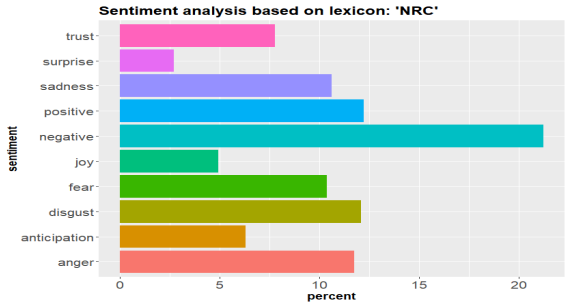

Figura 6.28: Sentimiento dominación

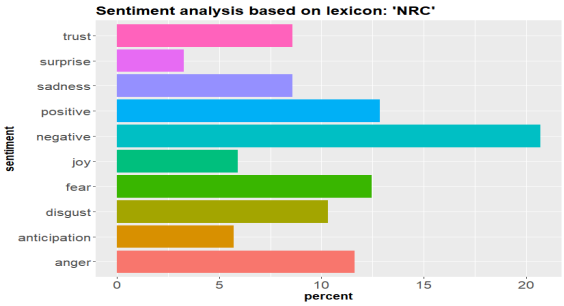

Figura 6.29: Sentimiento estereotipo

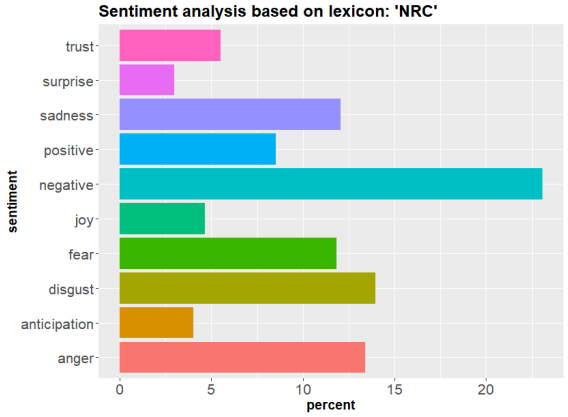

Figura 6.30: Sentimiento ataque sexual

En las gráficas de sentimiento no vemos diferencias pronunciadas entre los distintos tipos de misoginia, y la distribución que siguen es similar a la vista en la figura 6.25 al obtener la gráfica de sentimiento para los tweets misóginos en su conjunto.

Para finalizar este apartado, vemos a continuación, las gráficas de sentimiento para los distintos objetivos de la misoginia:

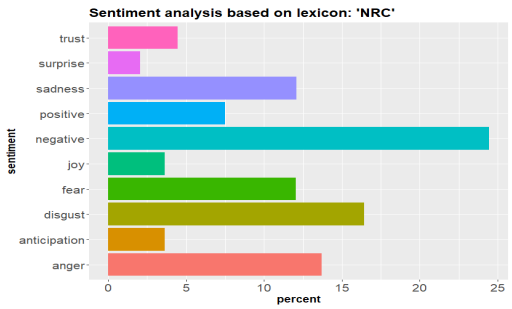

Figura 6.31: Sentimiento activo

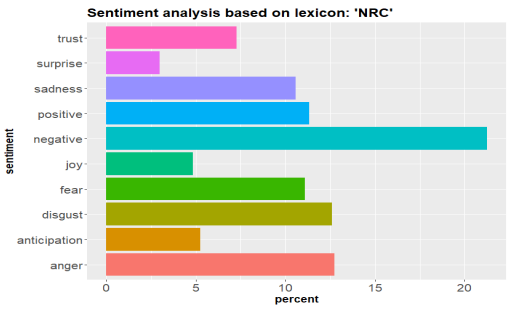

Figura 6.32: Sentimiento pasivo

Los resultados obtenidos con estas gráficas se sitúan en la línea de los ejemplos anteriores, donde vemos que el principal estado que reflejan los tweets misóginos es la negatividad, con sentimientos predominantes de miedo, disgusto y enfado.

Como se ha comentado anteriormente, el sentimiento de los tweets no se ha tenido en cuenta como una variable para entrenar nuestro modelo de detección de la misoginia. El principal motivo de esta decisión se debe a que los sentimientos que se han obtenido con léxico NRC, no son representativos de un comportamiento misógino exclusivamente, sino que pueden pertenecer a otras áreas. Por ejemplo, los sentimientos predominantes en los tweets misóginos son negatividad, miedo, disgusto y enfado, los cuales pueden estar presentes en otros comportamientos diferentes a la misoginia, tales como el racismo o la homofobia.

La utilización del sentimiento como una variable para entrenar un modelo de *machine learning* se podría aplicar en el caso de la detección del *hate speech*, ya que se haría referencia al discurso del odio en general, puesto que los sentimientos que hemos obtenido con el léxico NRC nos permitirían identificar el discurso del odio de manera genérica, pero no diferenciar entre sus diferentes categorías al utilizar un léxico similar con sentimientos comunes.

Una vez vistas las estadísticas obtenidas durante el desarrollo de nuestro algoritmo, pasamos a comprobar los resultados obtenidos en la evaluación de nuestro sistema de detección de la misoginia y clasificación de la misma.

6.2.3. Detección de la misoginia

En este apartado se reflejan los resultados finales que se han obtenido con nuestro sistema basándonos en las métricas de evaluación explicadas en el apartado 6.1.2. A continuación, se muestran los resultados obtenidos en función de la subtarea realizada.

En primer lugar, para la subtarea de la detección de la misoginia, se ha añadido la precisión resultante de nuestro algoritmo de aprendizaje automático sin distinguir entre la precisión de cada una de las clases, es decir, calculando la precisión media de nuestro algoritmo. Se refleja en la siguiente tabla:

SISTEMA	PRECISIÓN
SVM KERNEL LINEAL	0.7096
SVM KERNEL LINEAL CROSS-VALIDATION C	0.7312
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.7472

Tabla 6.7: Precisión de acierto en tweets misóginos

Es necesario tener en cuenta, que tal y como se ha explicado en el capítulo 4, una vez entrenado nuestro algoritmo en el *dataset* de *training*, se probará su eficacia sobre los datos de test. Inicialmente, aplicaríamos nuestra función SVM sin tunning, posteriormente, realizaríamos *cross-validation* con un kernel lineal para el parámetro de tunning C ("*cost*") y obtendríamos que el modelo con una mayor precisión tendría un valor de C

de 0.1, obteniendo una mejora en la precisión del modelo. Finalmente, para obtener una mayor precisión, utilizaremos un kernel radial y *cross-validation* para los parámetros de C y gamma (γ), cuyos valores más eficaces serían $C = 1$ y gamma (γ) = 0.5.

Calculamos el resto de métricas de evaluación definidas en el apartado 6.1.2 y los resultados son los siguientes:

Precisión:

SISTEMA	MISÓGINO	NO MISÓGINO
SVM KERNEL LINEAL	0.7041	0.7541
SVM KERNEL LINEAL CROSS-VALIDATION C	0.7057	0.7556
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.7271	0.7711

Tabla 6.8: Precisión de acierto en detección de misoginia

Recall:

SISTEMA	MISÓGINO	NO MISÓGINO
SVM KERNEL LINEAL	0.8388	0.8799
SVM KERNEL LINEAL CROSS-VALIDATION C	0.8838	0.9012
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.8156	0.9389

Tabla 6.9: *Recall* en detección de misoginia

Métrica F_1:

SISTEMA	MISÓGINO	NO MISÓGINO
SVM KERNEL LINEAL	0.7737	0.8023
SVM KERNEL LINEAL CROSS-VALIDATION C	0.7837	0.8122
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.7805	0.8240

Tabla 6.10: Métrica F_1 en detección de misoginia

Para la subtarea 1, se puede comprobar que, con la utilización de un kernel radial, la precisión de nuestro modelo mejora con respecto a los resultados iniciales. Se obtiene prácticamente un 75 % de efectividad a la hora de predecir tweets misóginos, siendo una precisión elevada y que podemos considerar como buena.

En el resto de métricas de esta tarea, se muestran sus resultados dependiendo de la correcta categorización de cada clase, en este caso, misógino, o no misógino. En general, para todas las métricas de evaluación definidas para esta tarea, obtenemos unos resultados generalmente buenos, ya que un buen modelo estadístico debe de tener las métricas de

precisión y F_1 con valores similares, y en este caso, esta premisa se cumple, además de obtener unos porcentajes altos.

La mejora que aporta la utilización del kernel radial se debe principalmente a dos factores: añadir una nueva dimensión para separar las clases y que gracias al parámetro gamma (γ), los márgenes de detección de clases no son tan estrictos como con un kernel lineal tradicional, tal y como se muestra en la figura 4.7.

A continuación, se reflejan los resultados obtenidos para la subtarea de clasificación de la misoginia según su tipología y objetivo.

Se muestran, en primer lugar, los resultados obtenidos para los sistemas de clasificación de la misoginia: desacreditar (DE), daño de imagen (DI), dominación (DO), estereotipo (ES) y ataque sexual (AS).

Precisión:

SISTEMA	DE	DI	DO	ES	AS
SVM KERNEL LINEAL	0.6797	0.6628	0.6044	0.7198	0.6267
SVM KERNEL LINEAL CROSS-VALIDATION C	0.6936	0.6069	0.6458	0.7126	0.7303
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.6024	0.5548	0.5601	0.6802	0.5646

Tabla 6.11: Precisión de acierto en categorías misóginas

Recall:

SISTEMA	DE	DI	DO	ES	AS
SVM KERNEL LINEAL	0.2864	0.3461	0.2352	0.4750	0.3333
SVM KERNEL LINEAL CROSS-VALIDATION C	0.2864	0.3461	0.2352	0.4750	0.3333
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.1982	0.1153	0.1323	0.375	0.1616

Tabla 6.12: *Recall* en categorías misóginas

Métrica F_1:

SISTEMA	DE	DI	DO	ES	AS
SVM KERNEL LINEAL	0.3785	0.4587	0.3298	0.5629	0.3905
SVM KERNEL LINEAL CROSS-VALIDATION C	0.3785	0.4587	0.3298	0.5629	0.3905
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.2249	0.1875	0.2168	0.5128	0.2461

Tabla 6.13: Métrica F_1 en categorías misóginas

Como podemos apreciar, se ha obtenido cada una de las métricas para cada categoría

misógina.

Para esta subtarea, vemos que el sistema con el que obtenemos mejores resultados es el SVM con un kernel lineal, independientemente del uso de *cross validation* con el parámetro de *Cost*, ya que el valor más óptimo para esta variable es 1, mismo valor que coge por defecto la función de kernel lineal sin *cross validation*. Con este parámetro, se busca la inclinación ideal de nuestra línea de separación de clases para separar con una mejor precisión las diferentes clases, como se ha representado en la figura 4.6. En este caso, C tiene un valor de 1, y es que cuanto más se aproxima a cero, menos se penalizan los errores, pero más observaciones pueden estar en el lado incorrecto.

Observando las tablas de las métricas de evaluación para esta subtarea, podemos apreciar como los resultados de precisión, exhaustividad y F_1 disminuyen considerablemente con respecto a los valores de la subtarea 1. De hecho, el mejor resultado de precisión media obtenido para esta subtarea ha sido de 0,6157 y si nos fijamos en las demás métricas, vemos como se reducen debido a la complejidad de la tarea, ya que es muy difícil clasificar un tweet en su categoría correcta de misoginia, de ahí que F_1 tenga un valor tan bajo, ya que no hay practicamente diferencia entre una categoría misógina y otra.

Por otro lado, se debe de resaltar el hecho de que obtengamos una precisión o exactitud bastante alta si la comparamos con las métricas *recall* y F_1. Esto es debido a que la precisión nos da la calidad de la predicción, es decir el porcentaje de clasificaciones correctas que hemos predicho de cada categoría y *recall* nos da la cantidad, es decir el porcentaje de clasificaciones correctas que hemos identificado. La métrica de F_1 está vinculada a estas dos medidas anteriores. Esta diferencia también aparece en algunas ocasiones debido a que las clases no están balanceadas, es decir hay mucha diferencia entre el número de clases de una categoría con respecto a otra. Por ejemplo, para este caso, vemos que muchos de los tweets misóginos están etiquetados dentro de la categoría desacreditar, y por el contrario, como daño de imagen solo están etiquetados 92 tweets, por lo que las clases no están muy bien balanceadas, ya que sería más fácil acertar diciendo que el tweet pertenece a la categoría desacreditar y esta casuística influye en la precisión de nuestro modelo causando la diferencia vista entre las métricas.

Para poder calcular la métrica Macro F_1(categoría de misoginia), utilizaremos el sistema SVM con kernel lineal, ya que es con el que mejores resultados hemos obtenido en esta subtarea. Calculando la media de esta métrica con los valores F_1 de las diferentes categorías misóginas, obtenemos un valor Macro F_1(categoría de misoginia) = 0.4240. De esta forma, ya tendríamos la primera métrica para calcular nuestro valor posterior de Macro F_1 en conjunto con el objetivo misógino.

Una vez obtenidas las métricas para la categorización de la misoginia, repetiríamos el proceso, pero esta vez, para el objetivo misógino, es decir, activo o pasivo. En esta subtarea, la situación sería similar a la detección de la misoginia, ya que tenemos simplemente dos clases. Los resultados para las métricas de evaluación de la clasificación de cada una de las clases son los siguientes:

Precisión:

SISTEMA	ACTIVO	PASIVO
SVM KERNEL LINEAL	0.7423	0.8126
SVM KERNEL LINEAL CROSS-VALIDATION C	0.8231	0.7621
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.8465	0.7890

Tabla 6.14: Precisión de acierto en objetivo de misoginia

Recall:

SISTEMA	ACTIVO	PASIVO
SVM KERNEL LINEAL	0.8547	0.901
SVM KERNEL LINEAL CROSS-VALIDATION C	0.8706	9201
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.8954	0.9393

Tabla 6.15: *Recall* en objetivo de misoginia

Métrica F_1:

SISTEMA	ACTIVO	PASIVO
SVM KERNEL LINEAL	0.8342	0.8867
SVM KERNEL LINEAL CROSS-VALIDATION C	0.8526	0.9102
SVM KERNEL RADIAL CROSS-VALIDATION C Y GAMMA (γ)	0.8659	0.9266

Tabla 6.16: Métrica F_1 en objetivo de misoginia

En este caso, al igual que en la subtarea 1, vemos que los mejores resultados los obtenemos con el uso de un kernel radial. Tanto la subtarea de la detección de la misoginia como la detección de su objetivo, tienen en común que están formadas por simplemente dos clases, en el primer caso, serían misógino y no misógino y en el segundo, activo y pasivo. El hecho de tener que buscar un algoritmo de SVM lo más preciso posible para la separación de dos clases, hace que el kernel radial sea el más óptimo para estos casos, donde también se ha utilizado *cross-validation* de los parámetros de C y gamma (γ).

La precisión media obtenida para este algoritmo de clasificación es de 77,9 %, siendo el el valor más alto obtenido en las tres subtareas. Igualmente, las métricas de F_1 y *recall* tienen unos valores altos, por lo que es una buena indicación de que nuestro sistema es válido.

Con la tabla 6.16, podemos obtener la métrica Macro F_1(objetivo) para nuestro sistema con una mayor precisión gracias al uso del kernel radial, y obtendríamos un valor medio de 0.8925.

Una vez obtenidas todas las métricas anteriores, estamos en disposición de conocer la métrica de evaluación de esta subtarea, tal y como se recoge en la ecuación 6.4. Esta métrica Macro F_1 tiene un valor de 0.6582.

Con respecto a las métricas de evaluación obtenidas en las diferentes subtareas, podemos afirmar que nuestros sistemas desarrollados para la detección de la misoginia y su clasificación en diferentes categorías y objetivos, tienen una buena eficacia.

Relativo a las tareas realizadas, podemos destacar la menor precisión que se ha obtenido en la clasificación de los tweets en diferentes categorías misóginas. Esto es debido, principalmente, a que es muy difícil, incluso para una persona, diferenciar una categoría misógina de otra, ya que el mismo tweet podría formar parte de más de un tipo de misoginia, por lo que la subjetividad juega un papel muy importante en la clasificación de nuestros tweets. Adicionalmente, el vocabulario que se utiliza en las diferentes categorías misóginas es muy similar, por lo que es muy complicado establecer un BoW único para cada categoría.

Para intentar reducir esta problemática, se podría utilizar una léxico de palabras violentas o sentimientos que estuvieran recogidos dentro de alguna de las categorías misóginas definidas y que fueran excluyentes, es decir que no se repitieran entre distintas tipologías, pero como ya se ha comentado, es complicado, ya que no existe una gran diferencia entre las distintas categorías misóginas establecidas a lo largo de este trabajo. A pesar de esto, aunque no lográsemos unos resultados con una precisión mucho mayor, nuestro sistema debería de experimentar alguna mejora.

Finalmente, para cerrar el capítulo, se destaca que los resultados obtenidos en las métricas de evaluación mostradas, se encontrarían dentro de los *baseline* establecidos en la tarea AMI (IBEREVAL, 2018).

7
Conclusiones y trabajo futuro

7.1. Conclusiones

Con la realización de este trabajo de fin de máster se ha analizado la problemática de la misoginia online presente hoy en día desde un punto de vista teórico y práctico.

No cabe duda de que el crecimiento exponencial de la web, ha aportado numerosos beneficios, tales como el acceso a gran cantidad de información, facilidad de uso en múltiples actividades cotidianas y la conexión entre diferentes usuarios. Sin embargo, la proliferación de la web, ha propiciado su utilización como una herramienta en la que la posibilidad de anonimato, la hace propensa para atacar a ciertos grupos sociales. Uno de los objetivos de estos ataques es el colectivo femenino, ya que, cada vez, observamos más episodios de acoso y *hate speech* en contra de la mujer, principalmente en las redes sociales. Este comportamiento se conoce como misoginia, y lo podemos definir como el odio o prejuicio hacia la mujer, pudiendo ser manifestada lingüísticamente en diversas formas, incluyendo exlusión social, discriminación, hostilidad, amenazas violentas y objetificación sexual.

Este comportamiento de misoginia online puede afectar a las mujeres en el mundo real de diversas formas, tales como trastornos psicológicos, depresión o incluso llegando algunas veces a desencadenar hechos más trágicos como es el suicidio. Surge, por lo tanto, la necesidad de frenar esta clase de comportamientos para poder evitar estas consecuencias, por lo que este trabajo de fin de master es un ejemplo de cómo la tecnología nos puede ayudar frente a un problema existente hoy en día en nuestra sociedad.

Con el análisis del estado del arte, se han repasado los diferentes estudios que se han realizado hasta el momento con respecto al *hate speech* y la misoginia online, y se ha visto cuáles han sido los enfoques que han utilizado otros autores para tratar la problemática a la que nos estamos enfrentando, al igual que las aproximaciones que han seguido.

En líneas generales, las aproximaciones utilizadas para la detección del *hate spech* y misoginia online son similares en los estudios realizados, y la mayoría se basan en la utilización de recursos léxicos como vocabulario específico del discurso del odio y vocabulario violento para la construcción de un *Bag of Words*, análisis de sentimientos, técnicas de *topic modeling* como puede ser LDA y figuras lingüísticas como los etiquetadores PoS, todo ello utilizado para desarrollar y entrenar algoritmos de aprendizaje automático basado en las técnicas de SVM, *Random Forest* o redes neuronales.

Para el desarrollo de nuestro sistema, hemos realizado en primer lugar, un procesamiento de texto, basado en las técnicas SASM, es decir, las técnicas de procesamiento enfocadas a redes sociales. Gracias a la ayuda de los recursos lingüísticos, hemos podido entender la taxonomía de los tweets misóginos para posteriormente diseñar y entrenar un modelo matemático que nos permita decidir de forma automática, si un tweet es misógino o no y realizar una clasificación de la misoginia en sus diferentes tipologías.

Como se ha mostrado, el algoritmo utilizado para el desarrollo del sistema ha sido SVM,

siendo una técnica de aprendizaje supervisado de *machine learning*.

Los resultados que se han obtenido son, de forma general, buenos, y se ha mejorado la precisión inicial obtenida gracias a la realización de iteraciones y *cross-validation* de los parámetros del algoritmo de SVM. Hemos visto que los mejores resultados los hemos obtenido en las tareas de detección de la misoginia y de su clasificación en el objetivo activo o pasivo, ya que es más fácil clasificar dos clases que estén bien balanceadas, y la utilización de un SVM con kernel radial nos ayuda a mejorar la precisión en estas tareas. Por otro lado, en la tarea de clasificación de la misoginia, se han obtenido peores resultados debido a que es más complicado establecer una clasificación en categorías que son muy parecidas entre sí, y no existe un vocabulario específico para cada una de ellas, por lo que un mismo tweet, podría pertenecer a más de una categoría.

Como puntos de mejora, se podría incluir la utilización de léxicos de vocabulario violento agrupado por diferentes categorías misóginas, siempre y cuando esté bien categorizado. También se podría hacer uso de un análisis de sentimiento de los tweets si los sentimientos están correctamente identificados en cada categoría misógina. Adicionalmente, en cuanto a los algoritmos de *machine learning*, se podrían combinar varios sistemas de aprendizaje supervisado como redes neuronales o *Random Forest* para intentar obtener una mejor precisión en nuestro algoritmo.

Este trabajo es un ejemplo de como el desarrollo de la tecnología, nos puede ayudar a solucionar problemas a los que nos enfrentamos diariamente. En relación a lo aprendido durante el Máster, me ha parecido un trabajo realmente interesante en el que se aplican conocimientos de diversas áreas y materias para solucionar un problema de la actualidad como es la misoginia online.

7.2. Trabajo futuro

Tal y como se ha defendido a lo largo de todo el trabajo, todavía quedan estudios por realizar para obtener un estado del arte completo.

Prácticamente todos los estudios que se han realizado acerca de la detección de la misoginia online se están basando en el análisis de texto, pero en la web, no solo encontramos este tipo de información, sino que también es necesario contar con los datos multimedia. Por este motivo, es necesario que los estudios y esfuerzos en la detección de la misoginia online, no solo vayan destinados al análisis de texto, sino más allá. Es necesario que aúnen fuerzas las técnicas de *machine learning*, con las llamadas técnicas y sistemas RIM (Recuperación de información multimedia).

Con la combinación de estas dos áreas del conocimiento, podremos ver en un futuro próximo, resultados prometedores en el campo de la detección de la misoginia online, no solo en textos, sino también en elementos multimedia, como pueden ser imágenes o audio.

Adicionalmente de la utilización de técnicas RIM, para trabajos futuros sería conve-

niente incluir información del usuario en el *dataset* que se analizara, ya que podríamos obtener una métrica de la tendencia de un usuario a publicar tweets misóginos. Igualmente, en la información del usuario se podría incluir la geolocalización, para comprobar si en ciertas áreas se publican más tweets misóginos que en otras.

Otra posible área de trabajo consistiría en analizar las interacciones que se producen en los tweets de los usuarios, ya que, como se ha comentado, la subjetividad juega un papel muy importante en la detección y clasificación de la misoginia, y en numerosas ocasiones, con un solo tweet, no somos capaces de asegurar con una precisión alta, si es misógino, pero analizando posibles comentarios de otros usuarios a ese tweet, podríamos reforzar el sistema para detectar si es misógino y su categoría.

Todavía queda mucho trabajo por realizar en materia de la detección de la misoginia online, pero no cabe duda de que se están realizando un mayor número de investigaciones, incorporando nuevas técnicas y metodologías y, al mismo tiempo, la sociedad está cada vez más concienciada de que se tiene que erradicar este problema.

Bibliografía

ACNUDH. https://www.ohchr.org/sp/issues/freedomopinion/articles19-20/, 2020.

M. Anzovino. Misogyny detection on social media: a methodological approach, 2018.

M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of mysogynistic language on twitter, 2018.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets, 2017.

J. Banks. Regulating hate speech online. international review of law, 2010.

Blei, Ng, and Jordan. Latent dirichlet allocation, 2003.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation, 2002.

P. Brown and S. Levinson. Politeness: Some universals in language usage, 1987.

P. Brown, P. Desouza, R. Mercer, V. Pietra, and J. Lai. Class-based n-gram models of natural language, 1992.

J. Canos. Misogyny identification through svm at ibereval 2018, 2018.

N. Chetty and S. Alathur. Hate speech review in the context of online social networks, 2018.

CRAN. https://cran.r-project.org, 2020.

CRAN-Caret. https://cran.r-project.org/web/packages/caret/caret.pdf, 2020.

CRAN-e1071. https://cran.r-project.org/web/packages/e1071/index.html, 2020.

CRAN-NLP. https://cran.r-project.org/web/packages/nlp/nlp.pdf, 2020.

CRAN-PKGraph. https://cran.r-project.org/src/contrib/archive/pkgraph/, 2020.

CRAN-Quanteda. https://cran.r-project.org/web/packages/quanteda/index.html, 2020.

CRAN-Textclean. https://cran.r-project.org/web/packages/textclean/index.html, 2020.

CRAN-TM. https://cran.r-project.org/web/packages/tm/tm.pdf, 2020.

CRAN-Topics. https://cran.r-project.org/web/packages/topicmodels/index.html, 2020.

J. Culpeper. Towards an anatomy of impoliteness, 1996.

F. Del Vigna, A. Cimino, F. Dell'Oretta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook, 2017.

N. Dennis, Z. Zuping, H. Damien, and J. Long. A lexicon-based approach for hate speech detection, 2015.

U. Dictionary. https://www.urbandictionary.com, 2020.

K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying, 2012.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings, 2015.

DUDH. https://www.un.org/es/documents/udhr/udhr-booklet-sp-web.pdf, 2020.

M. Duggan. Online harassment 2017, 2017.

Emolex. http://saifmohammad.com/webpages/nrc-emotion-lexicon.htm, 2016.

Facebook. https://www.facebook.com/communitystandards/hate-speech, 2020.

A. Founta, D. Chatzakou, N. Kourtellis, J. Balckburn, A. Vakali, and I. Leontiadis. A unified deep learning architecture for abuse detection, 2018.

S. Frenda, B. Ghanem, and M. Montes. Exploration of misogyny in spanish and english tweets, 2018.

J. García, M. Cánovas, R. Colombo, and R. Valencia. Detecting misogyny in spanish tweets. an approach based on linguistics features and words embeddings, 2020.

B. Ghanem, M. Montes, P. Rosso, and S. Frenda. Online hate speech against women: Automatic identification of misogyny and sexism on twitter, 2019.

GitHub-RDR. https://github.com/bnosac/rdrpostagger: :text=rdrpostagger2020.

Google. https://support.google.com/youtube/answer/2801939?hl=en, 2020.

C. Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions., 2010.

IBEREVAL. https://amiibereval2018.wordpress.com/, 2018.

H. Liu and P. Singh. Conceptnet—a practical commonsense reasoning tool-kit., 2004.

T. Lynn, P. Endo, P. Rosati, I. Silva, G. Santos, and D. Ging. A comparison of machine learning approaches for detecting misogynystic speech in urban dictionary, 2019.

S. Malmasi and M. Zampieri. Detecting hate speech in social media, 2017.

B. Mathew, D. Ritham, P. Goyal, and A. Murkherjee. Spread of hate speech in online social media, 2019.

S. Menini, G. Moretti, M. Corazza, E. Cabrio, S. Tonelli, and S. Villata. A system to monitor cyberbullyng based on message classification and social network analysis., 2019.

ML-Mastery. https://machinelearningmastery.com/gentle-introduction-bag-words-model/: :text=a2019.

E. Pamungkas, A. Cignarella, and V. Patti. 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets, 2018.

E. Periódico. https://www.elperiodico.com/es/sucesos-y-tribunales/20200524/jueza-cierra-caso-iveco-suicidio-videos-sexuales-trabajo-7971920, 2020.

G. Pitsillis, H. Ramampiaro, and H. Langseth. Detecting ofensive language in tweets using deep learning, 2018.

M. Plaza, D. Molina, A. Urena, and T. Martin. Detecting misogyny and xenophobia in spanish tweets using language technologies, 2020.

B. Poland. Haters: Harassment, abuse, and violence online, 2016.

PRC. Online harrassment 2017, July 2017.

RDocumentation-Methods. https://www.rdocumentation.org/packages/methods/versions/3.6.2, 2020.

A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing, 2017.

ScienceDirect. https://www.sciencedirect.com/topics/mathematics/document-matrix, 2013.

E. Shushkevich and J. Cardiff. Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018, 2018.

Significados. https://www.significados.com/misoginia/, 2020.

L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media, 2016.

J. Suler. The online disinhibition effect., 2004.

Twitter. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, 2020.

Z. Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, 2016.

Wikipedia. https://en.wikipedia.org/wiki/convention-on-the-elimination-of-all-forms-of-discrimination-against-women, 2020a.

Wikipedia. https://en.wikipedia.org/wiki/genocide-convention, 2020b.

Wikipedia. https://en.wikipedia.org/wiki/international-convention-on-the-elimination-of-all-forms-of-racial-discrimination, 2020c.

Wikipedia. https://en.wikipedia.org/wiki/convention-on-the-elimination-of-all-forms-of-discrimination-against-women, 2020d.

Wikipedia. https://en.wikipedia.org/wiki/international-covenant-on-civil-and-political-rights, 2020e.

Wikipedia. https://es.wikipedia.org/wiki/n-grama, 2020f.

Wikipedia. https://en.wikipedia.org/wiki/vladimir-vapnik, 2020g.

Wikipedia. https://es.wikipedia.org/wiki/misoginia, 2020h.

WMC. https://www.womensmediacenter.com/profile/bailey-poland, 2020.

F. Yang, X. Peng, G. Ghosh, R. Shilon, H. Ma, E. Moore, and G. Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification, 2019.