UNIVERSIDAD DE CASTILLA-LA MANCHA

ESCUELA SUPERIOR DE INFORMÁTICA

GRADO EN INGENIERÍA INFORMÁTICA

Computación

TRABAJO FIN DE GRADO

Análisis de Sentimientos para la prevención de mensajes de odio en las Redes Sociales

Andrés Montoro Montarroso

Febrero, 2019

UNIVERSIDAD DE CASTILLA-LA MANCHA
ESCUELA SUPERIOR DE INFORMÁTICA

Tecnologías y Sistemas de Información
Derecho Público y de la Empresa

Computación

TRABAJO FIN DE GRADO

Análisis de Sentimientos para la prevención de mensajes de odio en las Redes Sociales

Autor(a): Andrés Montoro Montarroso

Director(a): Dr. José Ángel Olivas Varela

Director(a): Dr. Adán Nieto Martín

Febrero, 2019

Análisis de Sentimientos para la prevención de mensajes de odio en las Redes Sociales
© Andrés Montoro Montarroso, 2019

Este documento se distribuye con licencia Creative Commons Atribución Compartir Igual 4.0. El texto completo de la licencia puede obtenerse en https://creativecommons.org/licenses/by-sa/4.0/. El escudo de Informática utilizado por este documento ha sido realizado por Franciso Moya, David Villa e Ignacio Díez, su inclusión debe respetar los derechos de autor y las licencias a las que se vea sometido. La copia y distribución de esta obra está permitida en todo el mundo, sin regalías y por cualquier medio, siempre que esta nota sea preservada. Se concede permiso para copiar y distribuir traducciones de este libro desde el español original a otro idioma, siempre que la traducción sea aprobada por el autor del libro y tanto el aviso de copyright como esta nota de permiso, sean preservados en todas las copias.

Tribunal:

Presidente: _____

Vocal: _____

Secretario: _____

Fecha de defensa: _____

Calificación: _____

Presidente	Vocal	Secretario
Fdo.:	Fdo.:	Fdo.:

A Isabel, José Andrés y
Antonio

Resumen

Internet a modificado la forma de comunicación en sociedad. Con la aparición de las Redes Sociales se ha desarrollado un fenómeno que ha supuesto una mayor intercomunicación entre usuarios y ha transformado Internet en un extraordinario vehículo para la difusión de mensajes. Esto ha expuesto un nuevo universo para la difusión de Delitos de Odio que no preexistía.

La comunicación ofensiva y de incitación a la violencia que pueda constituir un daño contra un grupo o parte de él o incluso contra una persona determinada por su razón de pertenencia al mismo por motivos de etnia, género, orientación sexual, religión, grupo social, afiliación o ideologías políticas o por otras características sociales, personales o funcionales. Esto puede crear un clima de hostilidad y violencia cuya consecuencia puede llegar a poner en peligro a los nombrados colectivos.

El presente Trabajo fin de Grado (TFG) propone desarrollar un mecanismo computacional capaz de identificar y clasificar según su intensidad, mensajes de odio en las redes sociales utilizando técnicas de Análisis de Sentimientos, Procesamiento del Lenguaje Natural y Lógica Borrosa que partiendo de una taxonomía diseñada a partir de la legalidad vigente y el conocimiento de un experto permita determinar la intensidad del discurso de odio y las particularidades que lo componen para informar de ello y que se tomen las decisiones pertinentes atendiendo a la legalidad actual y a la responsabilidad social corporativa de cada compañía.

Abstract

Internet has changed the way of communication in society. With the emergence of Social Media, it has appeared a phenomenon that has meant a greater intercommunication between users, and has transformed Internet into an extraordinary vehicle for the dissemination of messages. This has exposed a new universe for the dissemination of Hate Crimes that did not exist before.

Offensive communication and incitement to violence that may constitute harm against a group or part of it or even against a person determined by reason of belonging to it on grounds of ethnicity, gender, sexual orientation, religion, social group, political affiliation or ideology or by other social, personal or functional characteristics. This can create a climate of hostility and violence whose consequence may endanger the named social groups.

This project proposes developing a computational mechanism capable of identifying and classifying according to their intensity, hate messages in social media using techniques of Sentiment Analysis, Natural Language Processing and Fuzzy Logic. The starting point is a taxonomy designed from the current legality and the knowledge of an expert allows to determine the intensity of the hate speech and the particularities that compose it to inform it and that the pertinent decisions are taken considering the prevalent legality and the corporate social responsibility of each company.

AGRADECIMIENTOS

A lo largo de mi vida he querido ser muchas cosas, desde paleontólogo, pasando por conductor de tren, hasta ingeniero aeroespacial. No fue hasta mi último año de instituto que decidí estudiar Ingeniería Informática y no puedo estar más orgulloso de mi elección. En esta última etapa de este maratón que ha sido la carrera no puedo dejar de agradecer a todas las personas que me han acompañado durante estos años.

En primer lugar a mi madre, Isabel, es difícil expresar con palabras lo agradecido que me siento. Es un ejemplo de vida para mí, una mujer fuerte, trabajadora, inteligente capaz de luchar por su familia lo indecible. Ha sido, es y será mi apoyo más fundamental.

A Toñín, mi hermano, la mejor persona que conozco, un pilar fundamental en mi vida. Qué aburridos serían los días sin nuestras discusiones y berridos.

A mi padre, José Andrés. Un ejemplo de esfuerzo y dedicación me ha enseñado a ser la persona que soy ahora.

A mis abuelas, Maruja y Julia que siempre han estado ahí para ayudarme.

A mis tías, Juli, Toñi y María (la tata) que me han querido como a un hijo (en caso de la tata como a un hermano).

A Cristian, mi mejor amigo que desde hace más de 9 años me acompaña compartiendo aficiones y aventuras. Sin su apoyo este trabajo no hubiera sido posible.

A mis amigos los de siempre Antonio, Álvaro, Alberto, Roberto, Ángel, Carlos, María y Marino por hacerme tener la mejor infancia y adolescencia del mundo.

A mis compañeros de Prometeus que me han ayudado a crecer como profesional. En especial a Josevi, que empezó siendo compañero y ha acabado siendo un gran amigo, gracias por enseñarme tanto. Y a Eloy, que con sus estornudos y chistes escatológicos los días en la oficina se hacían más amenos.

Por último, y no menos importante a mis directores. José Ángel por el que siento gran admiración, agradecerte todo el apoyo y la confianza que me has brindado durante toda mi etapa universitaria. Adán amigo mío y de la familia, es todo un orgullo y un privilegio trabajar contigo.

En definitiva, gracias a todas las personas que me habéis acompañado durante esta bonita etapa.

Andrés Montoro Montarroso

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE LISTADOS

CAPÍTULO 1
INTRODUCCIÓN

Desde la aparición de la Web 2.0, y dentro de este marco, la aparición de las Redes Sociales (RRSS), se ha desarrollado un nuevo ámbito de comunicación que no preexistía. Esto ha facilitado la difusión masiva de mensajes a través del ciberespacio y consecuentemente su ratio de alcance. Este fenómeno ha supuesto una mayor intercomunicación entre usuarios y ha transformado Internet en un extraordinario vehículo para la difusión de mensajes [15]

Internet ha modificado las condiciones de comunicación en sociedad y se ha convertido, entre otras cosas, por sus características de neutralidad, ausencia de censuras y por su constante desarrollo, en un nuevo ámbito de oportunidad delictiva distinto al del mundo físico [51]. Esto ha dado lugar a una mayor difusión de los delitos de odio y por consiguiente un mayor efecto.

Las formas de comunicación ofensiva y de incitación a la violencia que pueda constituir un daño contra un grupo o parte de él o incluso contra una persona determinada por su razón de pertenencia al mismo por motivos de etnia, género, orientación sexual, religión, grupo social, afiliación o ideologías políticas o por otras características sociales, personales o funcionales se remonta a los orígenes de las sociedades pero no fue hasta después de la Guerra de Secesión en Estados Unidos donde se empezaron a proteger a los esclavos recién liberados y a otros colectivos de la discriminación y la violencia [42],y en Europa, hasta después de la Segunda Guerra Mundial ulteriormente de la amarga experiencia del fascismo y el nazismo y de la creciente amenaza del estalinismo en el este de Europa, donde a partir del Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales de 1950 (CEDH), surgieron las primeras regulaciones jurídicas sobre el Discurso del Odio. En España, ya se habla de los derechos fundamentales de cada uno de los Españoles independientemente de su nacimiento, raza, sexo, religión, opinión o cualquier otra condición o circunstancia social en el artículo 14 de la Constitución Española[1]. Pero no fue hasta la entrada en vigor del vigente Código Penal (CP en adelante) de 1995 que la normativa penal antixenófoba con vocación de tutela antidiscriminatoria y del principio de igualdad, experimentó figuras penales más relevantes cómo una agravante genérica que castiga la comisión de cualquier delito con pretexto racista, ideológico, sexo, orientación o identidad sexual, etcétera (art. 22.4º CP ver Anexo A); un precepto para castigar la comisión de delitos de incitación e injurias por las razones anteriores en el artículo 510 del CP (ver Anexo A) pasando también por el hecho de denegar por razones discriminatorias las prestaciones de servicios públicos o en el propio ejercicio de actividades profesionales (artículo 511 y 512 del CP), etcétera[40]

Las Redes Sociales, foros en Internet, secciones de comentarios en los distintos medios

[1] http://www.congreso.es/consti/constitucion/indice/titulos/articulos.jsp?ini=14&tipo=2

de comunicación de prensa digital, han contribuido a la universalización del discurso de odio gracias al anonimato que proporcionan y a la cantidad de personas a las que este discurso llega.

Dada la inmensa cantidad de mensajes que acaecen a diario en los distintos medios de "*Social Media*", se hace imposible controlar el contenido de forma manual, por ejemplo en medios de comunicación como EL PAIS se producen 13.000 comentarios al día, pero si observamos las grandes RRSS la cifra se vuelve astronómica, en Twitter se publican una media de 500 millones de tweets al día. Existen diversas formas de moderar estos comentarios ilícitos, tanto de forma manual en foros o secciones de comentarios con moderadores,como de forma automática empleando diversas técnicas para ello, como el Machine Learning o bolsas de términos y por los propios usuarios, práctica cada vez más habitual ya que las grandes Redes Sociales proporcionan a sus usuarios mecanismos para denunciar contenido inapropiado que posiblemente conllevará la acción de eliminar ese mensaje o incluso una posible consecuencia penal. Aún así este tipo de denuncias se convierte en intratable y el principal problema es confiar en el usuario la capacidad de juzgar un mensaje por su contenido, ya que el usuario se entiende que no es jurista y no está capacitado para clasificar este tipo de mensajes por lo que la mayoría de denuncias podrían ser falsas.

Como se ha expuesto el tema que aborda el Trabajo de Fin de Grado es el diseño de un mecanismo capaz de detectar posibles discursos de odio en las Redes Sociales con el fin de prevenir este tipo de comunicación violenta y facilitar al medio o red social la tarea de localizarlos y actuar en consecuencia. Para ello se emplearán técnicas de Inteligencia Artificial (IA en adelante)[65], concretamente se hará uso de técnicas de Procesamiento de Lenguaje Natural (PLN en adelante), que es un área de investigación que tiene como objetivo la tarea de procesar cómo los humanos entendemos y utilizamos el lenguaje para poder desarrollar herramientas que sean capaces de "comprender" la forma de comunicación humana[21]. De entre todas las aplicaciones del PLN se hará uso de lo que se puede considerar como una subdisciplina o parte de la misma, conocida como Análisis de Sentimientos o Minería de Opinión [34] [43] [44] que es un campo de investigación que analiza sentimientos, valoraciones, actitudes y emociones hacia las entidades y sus atributos. Este nuevo campo de investigación ofrece muchas oportunidades para desarrollar nuevos servicios, especialmente debido al gran crecimiento de datos disponibles (RRSS, blogs, etc) y la necesidad de analizarlos. Por ejemplo, un sistema para reconocer malas opiniones sobre un producto o un servicio en una red social que analiza las publicaciones de sus usuarios para así detectar posibles comportamientos suicidas, etc. Dentro de la noción de Análisis de Sentimientos se pueden distinguir los siguientes sub-tareas que componen el concepto[36] :

- **Clasificación de Sentimientos**: consiste en analizar un documento para extraer una opinión sobre una entidad dentro del texto e intentar medir el sentimiento que tiene esa entidad dentro de dicho documento.
- **Clasificación de subjetividad**: es un mecanismo para detectar oraciones que se puedan interpretar de diferentes formas imitando el punto de vista humano al leer una oración, es decir, ser capaz de diferenciar información objetiva de la que expresa sentimientos, emociones, opiniones, etc.
- **Resumen de opinión**: está especialmente focalizado en extraer las características más importantes de un texto que te permitan conocer el contexto y contenido del mismo.

- **Recuperación de opinión**: esta técnica intenta recuperar documentos que expresan una opinión sobre una consulta determinada.
- **Sarcasmo e ironía**: centrado en la búsqueda de sarcasmo e ironía dentro de un documento. La complejidad de esta utilidad es alta debido a la dificultad de formalizar este tipo de expresiones burlonas y mordaces.

Todo el proceso de Análisis de Sentimientos se apoyará en una **representación borrosa**[**zadeh1996fuzzy**] de los mismos para ayudar a modelar de manera precisa las ambigüedades del lenguaje y con ayuda de un experto y la literatura para obtener un enfoque basado en el léxico, la sintaxis y la semántica, permitirá cubrir ampliamente el espacio de entrada y salida para estudiar que características tiene un mensaje para ser considerado un posible delito de odio.

De entre todos los delitos de odio, contenidos en el artículo 510 del CP (ver Anexo A), tienen especial interés para el modelado del sistema los siguientes preceptos:

1. Serán castigados con una pena de prisión de uno a cuatro años y multa de seis a doce meses:

    a) "Quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, la hostilidad, discriminación o violencia, contra un grupo, una parte del mismo o contra una persona determinada por razón de su pertenencia a aquél, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual por razones de género, enfermedad o discapacidad".

2. Serán castigados con la pena de prisión de seis meses a dos años y multa de seis a doce meses:

    b) "Quienes lesionen la dignidad de las personas mediante acciones que entrañen humillación, menosprecio o descrédito de alguno de los grupos a que se refiere el apartado anterior, o de una parte de los mismos, o de cualquier persona determinada por razón de su pertenencia a ellos por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad, o produzcan, elaboren, posean con la finalidad de distribuir, faciliten a terceras personas el acceso, distribuyan, difundan o vendan escritos o cualquier otra clase de material o soportes que por su contenido sean idóneos para lesionar la dignidad de las personas por representar una grave humillación, menosprecio o descrédito de alguno de los grupos mencionados, de una parte de ellos, o de cualquier persona determinada por razón de su pertenencia a los mismos".

Debido a la ambigüedad del llamado ***Delito de Odio*** la jurisprudencia española no se pone de acuerdo en cómo juzgar este tipo de actividad penal. La herramienta no pretende actuar ni sustituir en ningún momento la figura del juez, con lo cuál acuñaremos el término de ***Comunicación Violenta y de Odio*** (CVydO en adelante)[52] y ***Discurso de Odio*** indistintamente para evitar esta ambigüedad.

De entre los diversos colectivos o grupos sociales que puedan resultar amparados por el artículo 510 del CP, el sistema de centrará en la detección de los mensajes de Comunicación Violenta y de Odio dirigidos contra la población Árabe que serán referidos como ***Colectivo Diana***. Este proyecto representa una necesidad actual, dado los terribles acontecimientos que azotan el mundo perpetrado por terroristas que actúan en nombre del Islam, está incrementando el racismo y la xenofobia a las personas de origen árabe que nada tienen que ver con el terrorismo, por tanto, es necesario evitar tanto como sea posible que se expanda este odio irracional.

1.1 JUSTIFICACIÓN

Es incontestable que los conocidos como Delitos de Odio han venido para quedarse y solo es necesario observar la hemeroteca recopilada por Jon-Mirena Landa Gorostiza en [38] para tener cuenta de ello:

"Quema simbólica del Coran en EEUU por un pastor causa revueltas y muertos en Afganistán durante varios día (EL PAIS, 3/4/2011"; "El Supremo revisa la condeno a una librería de Barcelona (Kalki) por vender material nazi (EL PAIS, 30/3/2011)"; "La campaña del PP, «Primero, los de Bilbao», denunciada por SOS Racismo (EL PAIS, 8/5/2012)"; "La muerte de un nigeriano hace estallar la violencia racial en Palma. Amigos del fallecido acusan a varios gitanos de haberle tirado desde un cuarto piso (EL PAIS, 30/8/2011)"; "La policía detiene al joven que lanzó un plátano a Dani Alves en El Madrigal (EL PAIS, 30/4/2014)"; "La comunidad judía denuncia a cinco tuiteros por mensajes antisemitas (EL PAIS, 20/5/2014)"; "Agredido un joven de 16 años por un grupo de neonazis en un bar de Manresa (El PAIS, 26/3/2012)"; "Archivada la querella contra Albiol por los panfletos que vinculaban rumanos con delincuencia (EL MUNDO, 4/6/2012)"; "Un jurado federal de Ohio condena a 16 amish por crímenes motivados por el odio. Los acusados cortaron el pelo y la barba de otros miembros de su propia comunidad. Estos hechos son considerados una humillación bajo las leyes que rigen su fe (EL PAIS, 21/9/2012)"; "La Fiscalía obre diligencias de investigación penal por las declaraciones de Sebastián sobre los gays. El arzobispo emérito de Iruñea: «Señalar a un gay una deficiencia no es una ofensa, es ayuda porque muchos casos se pueden normalizar con un tratamiento adecuado» (DEIA, 5/2/2014)"; "Multan con 4.000 euros a un aficionado «rojillo» por proferir insultos xenófobos; gritó «judío cabrón» y «judío muérete» a Aouate. Además, la Comisión Antiviolencia también le ha prohibido el acceso a los recintos deportivos por un periodo de un año (DEIA, 10/2/2011)"; "La Ertzainza investiga una agresión homófoba en un bar de Algorta (EL CORREO, 8/4/2015)"; "Detenido un grupo musical por incitar al odio contra los discapacitados (EL PAIS, 26/4/2015)"; "Marine le Pen juzgada por incitar al odio contra los musulmanes (EL PAIS, 20/10/2015)"; "El fiscal tendrá que decidir si llamar «cerdos» a vascos y catalanes es delito (PUBLICO, 14/4/2015)"; "Imputan un delito de odio o la mujer de Eibar que deseó la muerte al niño con cáncer que quiere ser torero (EL CORREO, 3/12/2016)"; "Urkullu (Lehendakari vasco) dice que el ataque a lo mezquita (en Vitoria) produce «perplejidad y vergüenza» (EL CORREO, 16/3/2016)"; "El ministro del interior afirma que la agresión de Altsasu es un delito de odio (ABC, 18/10/2016)"; "Primera condena por catalonofobia o raíz de un tweet sobre el accidente de German wings (EL MUNDO, 16/3/2017)"; "La Audiencia de Madrid admite la denuncia de una asociación franquista contra Wyoming y Dani Mateo (PUBLICO, 5/4/2017)"; "Antiviolencia llevará a la fiscalía la pitada al himno en la final de Copa (El PAIS, 1/6/2015)"; "Jean Marie Le Pen dice que las cámaras de gas

son «un detalle» histórico (PUBLICO, 3/4/2015)"; "La Fiscalía investiga posibles delitos de odio, amenazas y coacciones por la expulsión de fuerzas de Seguridad de hoteles de Cataluña (EL MUNDO, 3/10/2017)"; "La Asamblea Constituyente de Venezuela (ANC) aprueba una Ley contra el Odio, por lo Convivencia Pacífica y la Tolerancia (PANORAMA, 8/11/2017)"

Sólo en el año 2016 se registraron 1272 incidentes en España motivados por el odio en el **Informe sobre la evolución de los incidentes relacionados con los Delitos de Odio en España**[2] emitido por Ministerio del Interior.

Queda expuesta la necesidad de establecer mecanismos que eviten y prevengan este tipo de delitos y ayuden a su inmediata detección.

Otro punto muy importante a la hora de justificar el proyecto es en el marco de la responsabilidad de los proveedores de servicio de Internet[31], más concretamente de los prestadores de servicios de alojamiento o almacenamiento de datos, que abarcaría todo el espectro de "*Social Media*" que almacenen y distribuyan comentarios. En la Ley 34/2002, de 11 de julio, de servicios de la sociedad de la información y de comercio electrónico[3], en su Artículo 16 (Ver Anexo B) exime de responsabilidad a los proveedores de este servicio en los siguientes casos:

1. Si el contenido no ha sido distribuido por el propio proveedor (empresa proveedora).

2. No haya controlado o decidido sobre la publicación, es decir, cualquier Red Social o foro en Internet moderado está fuera del amparo de esta ley.

3. Si existe ese contenido ilícito solo podrán ser responsables si el órgano competente ha declarado la ilicitud de los datos, donde el proveedor deberá retirar o imposibilitar el acceso a ese contenido de manera diligente, es decir, con un tiempo suficiente para garantizar que la eliminación o denegación del acceso a esos datos no vulnere ningún derecho fundamental del resto de clientes del servicio.

Dado que es común que las empresas de RRSS tengan políticas de responsabilidad social corporativa y se dediquen a perseguir contenido ofensivo según sus políticas, esta ley no les ampararía por incumplimiento del punto 2 del artículo 16 de la Ley de Servicios de la Sociedad de la Información y del comercio electrónico (LSSI[4]) y por tanto, si existe contenido penalmente relevante se les aplicaría el Artículo 31 del CP[5]. Y debido a la importancia del control del Delitos de Odio en las RRSS, el pasado 1 de enero de 2018, entró en vigor la Ley "Netzwerkdurchsetzungsgesetz" más conocida como Ley **NetzDG**[9] donde establece un plazo de 24 horas a partir de la recepción de la denuncia de contenido ilícito para suprimir o bloquear dicho contenido y multa de hasta 5 millones de euros en caso de no cumplir la ley, entre otras regulaciones sobre contenido de odio. Esta ley se aplicará a Redes Sociales con más de 2 millones de usuarios registrados Alemanes (Twitter, Facebook, Youtube, etcétera), por tanto la necesidad de detectar este tipo de delitos es innegable.

[2] http://www.interior.gob.es/web/servicios-al-ciudadano/delitos-de-odio/estadisticas

[3] https://www.boe.es/buscar/act.php?id=BOE-A-2002-13758

[4] https://www.boe.es/buscar/act.php?id=BOE-A-2002-13758

[5] El que actúe como administrador de hecho o de derecho de una persona jurídica, o en nombre o representación legal o voluntaria de otro, responderá personalmente, aunque no concurran en él las condiciones, cualidades o relaciones que la correspondiente figura de delito requiera para poder ser sujeto activo del mismo, si tales circunstancias se dan en la entidad o persona en cuyo nombre o representación obre.

1.2 ESTRUCTURA DEL DOCUMENTO

En este apartado se pretende dar al lector una idea general del contenido del presente documento, donde se explicará de forma breve cada uno de los capítulos del mismo, de modo que su accesibilidad a información de su interés sea lo más sencilla posible.

Capítulo 1: Introducción

Donde se establece una visión global del proyecto a abordar y la justificación del mismo.

Capítulo 2: Objetivo

Propósito y alcance del proyecto.

Capítulo 3: Estudio de Viabilidad

Análisis exhaustivo de la viabilidad del proyecto a realizar.

Capítulo 4: Estado del Arte

En este capitulo se presenta un estudio bibliográfico de los temas considerados en el proyecto.

Capítulo 5: Metodología

Capítulo en el que se indica el procedimiento metodológico para el desarrollo del proyecto.

Capítulo 6: Desarrollo

Capítulo principal del proyecto donde se ilustra el proceso de contextualización y desarrollo del mismo.

Capítulo 7: Evaluación y resultados

Dado un conjunto de prueba, se analiza los resultados obtenidos en la elaboración del trabajo.

Capítulo 8: Conclusiones

Por último, en este capítulo se valora el trabajo realizado con los objetivos del mismo y se establece el trabajo futuro que se desprende de la realización del presente proyecto.

CAPÍTULO 2
OBJETIVO

En este capítulo se pretende dar una visión precisa de los objetivos generales y específicos definidos para la realización del presente Trabajo Fin de Grado (TFG en adelante). A continuación serán descritos los objetivos nombrados de manera concreta para mostrar al lector una visión precisa del alcance del proyecto.

2.1 OBJETIVO GENERAL

El objetivo principal del proyecto es desarrollar un mecanismo computacional capaz de identificar y clasificar mensajes de CVydO atendiendo a una serie de criterios e informar de ello para que se tomen las decisiones pertinentes atendiendo a la legalidad vigente y a la Responsabilidad Social Corporativa de cada compañía. El trabajo quedará ilustrado mediante una aplicación web de la que se podrá extraer un informe del grado de intensidad de un CVydO. Para ello se llevarán a cabo los siguientes sub-objetivos.

2.2 SUB-OBJETIVOS

Para el alcance del objetivo principal del proyecto ha sido necesario realizar distintas tareas divididas en sub-objetivos. A continuación se expone un breve resumen de los mismos para ayudar al lector a comprender el flujo del proyecto.

2.2.1 Establecer el dominio del proyecto

Dada la complejidad de modelar la tipificación de Delito de Odio para detectar CVydO de todos los colectivos que ampara, se hizo necesario establecer un dominio concreto dentro de todos los colectivos recogidos en la redacción del Articulo 510 del CP (ver Anexo A).

De entre los diversos colectivos o grupos sociales que pueden resultar amparados por el Artículo 510 del CP el colectivo diana seleccionado para el proyecto será el colectivo "musulmán y/o árabe". Existen dos razones que generan actualmente un gran peligro de que este colectivo sea objeto de Delitos de Odio. En primer lugar, la inmigración ilegal, una parte significativa de los ciudadanos extranjeros que se encuentran en nuestro país, ya sea con residencia legal o no legal, pertenecen a este colectivo. El rechazo que en determinados sectores de la población genera la inmigración se centra en mensajes contra este colectivo.

En segundo lugar, y obviamente por razones de terrorismo. No es infrecuente que en el discurso de odio anti-islam se mezclen ambos argumentos.

2.2.2 Diseño de experimento para la obtención de términos del dominio

Se pretende diseñar un experimento en el que el dominio del mismo sea el nombrado en el apartado anterior para permitir identificar qué estructuras lingüísticas y términos componen esos mensajes de CVydO contra la comunidad musulmana en los distintos escenarios posibles, por ejemplo: "Escribe un comentario en el que animes a un indeterminado grupo de personas a que discriminen al colectivo musulmán en un determinado espacio (por ejemplo, en las escuelas)". El experimento debe cubrir en la medida de lo posible el espacio de entrada y salida del dominio, para que el resultado de la ontología que se extraerá a partir de sus términos sea lo más precisa posible abarcando todas la alternativas del lenguaje propio del dominio o argot para que la posterior detección de estos mensajes sea lo más rigurosa posible.

El resultado de este experimento también tendrá su utilidad en la validación del sistema computacional.

2.2.3 Creación de la ontología del dominio

Dado que la variedad morfo-sintáctica del experimento se puede considerar incompleta por distintos factores, ya sea por el propio sesgo que pueda ocasionar el experimento o por la escasez de riqueza lingüística que se puede encontrar en un mensaje dentro del marco de las RRSS. Se hace necesaria la construcción de una ontología del dominio en cuestión para conseguir un mecanismo computacional más completo y robusto.

También se hace necesario incluir términos propios de la tipificación del delito para una mejor clasificación de los CVydO.

2.2.4 Búsqueda de patrones para definir la taxonomía del Delito de Odio

Se establece un proceso exhaustivo de búsqueda de patrones que permitan identificar de manera precisa los CVydO. Con ayuda del experto y la literatura estableceremos qué parámetros intrínsecos componen un Delito de Odio para identificar mejor los mensajes de CVydO, con el objetivo de establecer un mecanismo computacional que sea capaz de identificar cuando es un mensaje de CVydO.

2.2.5 Definición del conjunto y etiquetas borrosas para el mecanismo de Análisis de Sentimientos

Dada la complejidad del sistema, una vez se han establecido los patrones que nos permiten detectar CVydO, es necesario construir el conjunto borroso a partir de ellos para modelar la

intensidad que tiene un mensaje de CVydO siguiendo tanto los parámetros intrínsecos que se pueden extraer del mensaje, su entorno y el clima de la sociedad actual. Una vez construido el modelo de representación borrosa de conocimiento, se asignará unas etiquetas lingüísticas con ayuda del experto para establecer el sentimiento de cada uno de los mensajes de CVydO. Todo esto dará como resultado un modelo borroso en el que cada mensaje tendrá un grado de pertenencia a cada una de las etiquetas lingüísticas que representen la intensidad agravante de cada uno de los mensaje de CVydO.

2.2.6 Implementación del modelo

Una vez establecido el mecanismo borroso de Análisis de Sentimientos para la clasificación de los mensajes de CVydO, se hace una traducción a código para tener una representación funcional del sistema computacional.

2.2.7 Despliegue e infraestructura

Por último, el servicio es desplegado con Docker[1], implementando una arquitectura para la construcción de una API REST con la que se pueda integrar el mecanismo computacional descrito y con una aplicación web en cliente desarrollada en Angular 7 para una interacción amigable con el sistema. El servicio también se ejecutará en la plataforma de computación en la nube Amazon Web Service[2] (en adelante AWS) para proporcionar un servicio accesible en remoto.

[1] https://www.docker.com
[2] https://aws.amazon.com/es/

CAPÍTULO 3

ESTUDIO DE VIABILIDAD

En este capítulo se va a analizar la viabilidad del proyecto para justificar su realización y la posible continuación del mismo. Dada la naturaleza del proyecto, al contar con un experto para la realización del sistema computacional, se ha decidido aplicar el estudio de viabilidad conocido como Test de Slagel[69].

3.1 TEST DE SLAGEL

El Test de Slagel es un estudio de viabilidad para Sistemas Expertos que se basa en la evaluación de una serie de características de forma numérica con una ponderación específica por cada una de ellas. El estudio de Viabilidad está compuesto por tres etapas:

1. **Definición de características.**
2. **Asignación de pesos a cada una de las características.**
3. **Evaluación de cada aplicación candidata.**

La definición de características se divide en cuatro dimensiones:

1. **Plausibilidad**: Dimensión con la que se determina si se cuenta con los medios necesarios para poder abordar la creación del proyecto. Para ello se analizan dos aspectos:
   - **Características del experto**: aspecto en el que se evalúa la aptitud del experto frente al problema que se plantea, evaluando su prestigio y su capacidad de articular sus métodos y procedimientos de trabajo. Y la actitud que tiene frente al proyecto propuesto.
   - **Características de la tarea que lleva a cabo el experto**: analiza el nivel de dificultad de la tarea asignada, teniendo en cuenta las habilidades que se requieren para su adecuada realización.
2. **Justificación**: Se analizan aspectos como la necesidad de la experiencia, es decir, el contexto que abarca la realización del proyecto, y la inversión a realizar donde se analizan los costes y el retorno de inversión y si existen soluciones alternativas.

3. **Adecuación**: se estudia si el proyecto planteado es adecuado para la metodología de realización de un Sistema Experto, o por el contrario es resoluble mediante el sentido común.

4. **Éxito**: se determinan a priori las posibilidades de que el sistema se realice de manera exitosa atendiendo a los recursos (humanos y materiales) estén disponibles, la capacidad de las personas implicadas en el desarrollo del proyecto, que la comunicación entre implicados sea lo más sencilla posible para un seguimiento correcto y avance positivo del proyecto, la utilidad del sistema y que se adecue a lo esperado por el experto.

Sobre cada una de estas cuatro dimensiones se establece una categoría de aplicación, donde identifica a quién está dirigida la tarea. Se puede discernir tres actores: el experto, el usuario o directivo o a la propia tarea. Cada una de las tareas especificadas en el test pueden ser esenciales o deseables, donde las primeras no pueden ser inferiores a 7. A cada una de las características se le asigna un peso entre 0 y 10 dependiendo de la importancia relativa de la misma.

Teniendo en cuenta todo lo anterior, el proceso de evaluación del proyecto es el siguiente:

1. Asignación de un valor a cada una de las características de cada una de las dimensiones. Este valor se comprende entre 0 y 10, que significan *ausente* o *totalmente presente* respectivamente. Hay que tener en cuenta que la característica esencial no puede alcanzar un valor menor de 7, si se diera el caso el cómputo total de la dimensión sería cero y la aplicación quedaría automáticamente rechazada.

2. Ponderación del valor de la característica respecto al peso de la misma.

3. Multiplicar para cada dimensión los valores ponderados de las características.

4. Extraer la media para cada dimensión de los valores ponderados de las características, calculando la raíz n-ésima del producto obtenido del apartado anterior empleando como índice el valor máximo de los índices usados en cada dimensión.

5. Por último, dividir la suma del resultado de cada dimensión entre 4 siendo el valor máximo posible 76,21.

A continuación se ilustra el cálculo de la viabilidad del presente TFG haciendo uso del Test de Slagel (para una mayor comprensión de las tablas mirar la leyenda de la Tabla 3.1):

Tabla 3.1: Leyenda.

Acrónimo	Significado	Rango
CAT	Categoría	
EX	Experto(s)	
TA	Tarea	
IDEN. CAR.	Identificador de la característica	
Pi	Identificador de la dimensión de Plausibilidad	P1 ... P10
Ji	Identificador de la dimensión de Justificación	J1 ... J7
Ai	Identificador de la dimensión de Adecuación	A1 ... A12
Ei	Identificador de la dimensión de Éxito	E1 ... E17
E	Esencial	0 ... 10[1]
D	Deseable	0 ... 10

3.1.1 Evaluación de la plausibilidad

- Asignación de valores para la obtención de la plausibilidad del proyecto (ver Tabla 3.2).
- Explicación de la valoración de las características.
  - P1: Cualquier especialista en derecho penal podría ser un experto adecuado, dado su conocimiento para la interpretación de las leyes. Por tanto existen múltiples expertos.
  - P2: El experto Adán Nieto Martín es catedrático de derecho penal de la Universidad de Castilla-La Mancha. Sus principales líneas de investigación han sido el derecho penal económico y la construcción del derecho penal en la Unión Europea. En los últimos años ha participado y dirigido proyectos de investigación sobre responsabilidad penal de las personas jurídicas, cumplimiento normativo y, más recientemente, sobre prevención de la corrupción en administraciones públicas. Ha participado además en el asesoramiento a un gran número de empresas en la elaboración e implementación de programas de cumplimiento. Cuenta con numerosas obras destacadas en esta materia. Una lista completa puede consultarse en Dialnet [2]
  - P3: El experto es cooperativo por su gran interés por el proyecto y por la proximidad y accesibilidad para las entrevistas.
  - P4: Dado que su conocimiento del Código Penal es genuino y ha trabajado antes en Sistemas Expertos es capaz de articular sus métodos.
  - P5: Existen múltiples casos de prueba de CVydO, por ejemplo, en [52] hace un análisis de 250.836 extraidos de la Red Social Twitter días después de que se produjese el atentado en París contra los trabajadores del semanario satírico francés *Charlie Hebdo* en París, de los cuales el 2 % es considerado CVydO. También consideramos como casos de prueba los 259 mensajes extraídos del experimento realizado para el proyecto (Ver Anexo C).
  - P6: El proyecto está claramente especificado en el objetivo y acotado en un dominio concreto para hacer plausible su elaboración.

[2] https://dialnet.unirioja.es/

Tabla 3.2: Plausibilidad.

CAT.	IDEN. CAT.	PESO (P)	VALOR (V)	DENOMINACIÓN DE LA CARACTERÍSTICA	TIPO
EX	P1	10	**10**	Existen expertos.	E
EX	P2	10	**10**	El experto asignado es genuino.	E
EX	P3	8	**10**	El experto es cooperativo.	D
EX	P4	7	**10**	El experto es capaz de articular sus métodos pero no categoriza.	D
TA	P5	10	**8**	Existen suficientes casos de prueba; normales, típicos, ejemplares, correosos, etcétera.	E
TA	P6	10	**10**	La tarea está bien estructurada y se entiende.	D
TA	P7	10	**10**	Solo requiere habilidad cognoscitiva (no pericia física).	D
TA	P8	9	**9**	No solo se precisan resultados óptimos sino sólo Satisfactorios, sin comprometer el proyecto	D
TA	P9	9	**10**	La tarea no requiere sentido común.	D
DU	P10	7	**10**	Los directivos están verdaderamente comprometidos con el proyecto	D

  - P7: En ningún punto de la realización del presente TFG se requiere ningún tipo de pericia física.

  - P8: Dado que es una sistema computacional con un componente borroso y dentro de los juristas tampoco se ponen de acuerdo a la hora de delimitar un Delito de Odio no se requieren un 100 % de resultados óptimos. Aunque es deseable que todos los comentarios que se detecten sean mensajes de CVydO.

  - P9: El desarrollo del proyecto requiere un conocimiento profundo de Derecho Penal y de Inteligencia Artificial.

  - P10: Esta característica es ciertamente ambigua porque la naturaleza de este TFG no es empresarial, pero si de gran interés en el Instituto de Derecho Penal Europeo e Internacional[3].

- Evaluación de la dimensión de la aplicación candidata.

[3]http://institutoderechopenal.uclm.es/

$$VC_i = \prod_{j=1,2,5} (Vp_j//Vu)[\prod_{k=1}^{n} Pp_k \cdot Vp_k]^{1/n}$$
$$i = 1 \tag{3.1}$$
$$n = 10$$

donde:

- VC_i es el valor de la aplicación candidata en una dimensión (plausibilidad).
- Vp_j es el valor de la característica de plausibilidad en los tipos esenciales.
- Vu es el valor umbral del sistema.
- Pp_k peso de la característica de plausibilidad.
- Vp_k es el valor de la característica de plausibilidad en todos sus indicadores.

$$VC_1 = \mathbf{86{,}28} \tag{3.2}$$

3.1.2 Evaluación de la Justificación

- Asignación de valores para la obtención de la justificación del proyecto (ver Tabla 3.3).

Tabla 3.3: Justificación.

CAT.	IDEN. CAT.	PESO (P)	VALOR (V)	DENOMINACIÓN DE LA CARACTERÍSTICA	TIPO
EX	J1	10	**10**	El experto NO está disponible.	E
EX	J2	10	**9**	Hay escasez de experiencia humana.	D
TA	J3	8	**10**	Existe necesidad de experiencia simultánea en muchos lugares.	D
TA	J4	10	**9**	Necesidad de experiencia en entornos hostiles, penosos y/o poco gratificantes.	E
TA	J5	8	**9**	No existen soluciones alternativas admisibles.	E
DU	J6	7	**10**	Se espera una alta tasa de recuperación de la inversión	D
DU	J7	8	**10**	Resuelve una tarea útil y necesaria	E

- Explicación de la valoración de las características
  - J1: El experto está totalmente presente durante todo el proceso de desarrollo del proyecto.

- J2: Dado que el experto ha trabajado en sistemas expertos anteriormente y dado que no es la primera vez que surge una colaboración con el mismo se considera una experiencia alta.
- J3: El proyecto no requerirá necesariamente de juristas para su uso, es pensado principalmente para moderadores, pero totalmente factible para su uso en el ámbito juridico-profesional por su taxonomía del Delito de Odio.
- J4: No es necesaria la experiencia en entornos adversos.
- J5: No existen soluciones basadas en Inteligencia Artificial haciendo uso de lógica borrosa para detectar CVydO en España, existen guías para registro y monitorización (ver capítulo §4 para un estudio detallado del estado del arte).
- J6: Dado que se trata de un TFG la inversión económica ha sido mínima (sin tener en cuenta el hipotético gasto de personal en caso de tratarse de un proyecto remunerado) por tanto la tasa de recuperación es máxima.
- J7: Ver la sub-sección §1.1.

- Evaluación de la dimensión de la aplicación candidata

$$\begin{aligned} VC_i &= \prod_{j=1,4,5,7} (Vj_j // Vu)[\prod_{k=1}^{n} Pj_k \cdot Vj_k]^{1/n} \\ i &= 2 \\ n &= 7 \end{aligned} \tag{3.3}$$

donde:

- VC_i es el valor de la aplicación candidata en una dimensión (justificación).
- Vj_j es el valor de la característica de justificación en los tipos esenciales.
- Vu es el valor umbral del sistema.
- Pj_k peso de la característica de justificación.
- Vj_k es el valor de la característica de justificación en todos sus indicadores.

$$VC_2 = \mathbf{82{,}55} \tag{3.4}$$

3.1.3 Evaluación de la Adecuación

- Asignación de valores para la obtención de la adecuación del proyecto (ver Tabla 3.4).
- Explicación de la valoración de las características.
  - A1: En las reuniones con el experto se ha demostrado que las ideas que plantea el experto están perfectamente estructuradas.
  - A2: Dada la legalidad vigente y las políticas responsabilidad social corporativa de cada una de las compañías de RRSS, tiene un gran valor práctico.

- A3: El proyecto no está orientado a ningún tipo de estrategia empresaria, es un servicio que cubre necesidades actuales.
- A4: A parte de la universalización del Delito de Odio a través de las redes sociales y el gran problema actual con la comisión de este delito como se expresó en el apartado de introducción existe una nueva ley alemana llamada Ley NetzDG[9] que aumenta la responsabilidad de los proveedores de servicios, concretamente de RRSS donde la aplicación de este servicio sería de gran ayuda para los mismo. Y la legislación española no tardará en aplicar medidas similares.
- A5: El proyecto requiere de un alto conocimiento penalista y de Ingeniería del Conocimiento para extraer patrones para hacer uso de técnicas de IA.
- A6: El tamaño de TFG está perfectamente delimitado gracias al enfoque del proyecto en un solo Colectivo Diana (musulmán) para crear un prototipo inicial funcional y perfectamente escalable.
- A7: Gracias a la calidad del experto y a la ayuda de la literatura la transferencia de conocimiento es perfectamente plausible.
- A8: La penalización de este delito es extremadamente compleja ya que un sistema democrático es garante de libertad de expresión y excederse en la condena de este delito puede vulnerar este principio fundamental, por eso es de gran necesidad la taxonomía del delito con su posterior implementación en el proyecto para establecer un sistema de ayuda a la toma de decisiones eficaz y robusto.
- A9: No es de obligado cumplimiento que el resultado sea dado en tiempo real ya que el proceso de razonamiento es complejo pero es un objetivo deseable que pueda analizar mensajes en tiempo real[4].
- A10: La gran parte de conocimiento es proporcionada por el experto pero es deseable un conocimiento adquirido del dominio en cuestión.
- A11: El proceso de representación computacional de lenguaje natural es intrínseco a factores subjetivos.
- A12: El Código Penal es heurístico.

- Evaluación de la dimensión de la aplicación candidata.

$$\begin{aligned} VC_i = & \prod_{j=4,7,9,10} (Va_j//Vu)[\prod_{k=1}^{n} Pa_k \cdot Va_k]^{1/n} \\ & i = 3 \\ & n = 12 \end{aligned} \tag{3.5}$$

donde:

- VC_i es el valor de la aplicación candidata en una dimensión (adecuación).
- Va_j es el valor de la característica de adecuación en los tipos esenciales.
- Vu es el valor umbral del sistema.

[4] *Dada la capacidad computacional actual esta característica no debería ser esencial para la construcción de un sistema para la ayuda de toma de decisiones.*

- Pa_k peso de la característica de adecuación.
- Va_k es el valor de la característica de adecuación en todos sus indicadores.

$$VC_3 = \mathbf{58{,}64} \tag{3.6}$$

3.1.4 Evaluación del Éxito

- Asignación de valores para la obtención del éxito del proyecto (ver Tabla 3.5).
- Explicación de la valoración de las características
  - E1: Dado que es un proyecto común todos los implicados estamos en perfecta sintonía.
  - E2: Existe una larga trayectoria profesional por parte de los directores.
  - E3: Las reuniones aspiran a ser frecuentes con una perfecta sincronización con los implicados.
  - E4: Todas las sub-tareas tienen una gran repercusión en el resultado final del proyecto
  - E5: Al tratarse de un TFG existe un plazo de finalización, únicamente marcado por las convocatorias establecidas por la Universidad de Castilla-La Mancha. Podría decirse que el plazo es flexible ya que existen cuatro plazos de entrega por convocatoria y un total de tres convocatorias. La fecha de finalización del proyecto no implica la demora de otro proyecto.
  - E6: El proyecto consiste en la creación de un sistema computacional capaz de detectar CVydO en las redes sociales, por lo que gran parte del conocimiento está extraído de la legislación española. Dado que el proyecto será escalable atendiendo a las posibles modificaciones del precepto (ya que su eliminación se puede afirmar que es descartable) no es correcto decir que no depende de vaivenes políticos, pero estos no afectarían hasta el punto de hacer inservible el proyecto.
  - E7: Ver Capítulo §4.
  - E8: La metodología de desarrollo no seguirá el proceso de creación de un Sistema Experto de manera canónica.
  - E9: Cada paso será comprensible y documentado para una correcta comprensión del proyecto.
  - E10: Son perfectamente delimitables todas las tareas que conducen al éxito del proyecto.
  - E11: Se ha acotado el radio de acción de la persona jurídica que ampara el Artículo 510 del CP eligiendo un solo Colectivo Diana protegido por dicho artículo.
  - E12: La tecnología es totalmente válida y actual.
  - E13: No se trata de un sistema categórico si no uno de ayuda a la toma de decisiones de manera eficaz y efectiva en tiempo y calidad.

  - E14: De la herramienta se podrá exporta un documento de tipo PDF explicando el razonamiento seguido de manera clara y sencilla.
  - E15: La empresa podría considerarse el Instituto de Derecho Penal Europeo e Internacional[5] y el desarrollo del sistema no altera ningún proyecto en los que están embarcados.
  - E16: La realización de este TFG tiene proyección de futuro por parte de los implicados y un gran interés.
  - E17: Gracias a la documentación asociada al proyecto se garantiza una adecuada transferencia tecnológica.

- Evaluación de la dimensión de la aplicación candidata

$$\begin{aligned} VC_i &= \prod_{j=6,10,12,17} (Ve_j//Vu)[\prod_{k=1}^{n} Pe_k \cdot Ve_k]^{1/n} \\ n &= 4 \\ n &= 17 \end{aligned} \tag{3.7}$$

  donde:

  - VC_i es el valor de la aplicación candidata en una dimensión (éxito).
  - Ve_j es el valor de la característica de éxito en los tipos esenciales.
  - Vu es el valor umbral del sistema.
  - Pe_k peso de la característica de éxito.
  - Ve_k es el valor de la característica de éxito en todos sus indicadores.

$$VC_4 = \mathbf{59{,}44} \tag{3.8}$$

3.1.5 Evaluación de la viabilidad del sistema

Con el fin de evaluar el sistema, se procede al último caso en que que se calculará la media del sumatorio de todas los valores de las aplicaciones candidatas.

$$\begin{aligned} VC &= \sum_{i=1}^{4} VC_i/4 \\ VC &= 71,73 \end{aligned} \tag{3.9}$$

Una vez obtenido el valor de la aplicación candidata se calcula el porcentaje de viabilidad del proyecto:

[5] http://institutoderechopenal.uclm.es/

$$\begin{aligned} Viabilidad\,\% &= \frac{VC \cdot 100}{VCMax} \\ VCMax &= 76,21 \\ Viabilidad\,\% &= \mathbf{94{,}12\,\%} \end{aligned} \tag{3.10}$$

De acuerdo a los resultados de la evaluación el proyecto es altamente viable.

Tabla 3.4: Adecuación.

CAT.	IDEN. CAT.	PESO (P)	VALOR (V)	DENOMINACIÓN DE LA CARACTERÍSTICA	TIPO
EX	A1	5	**10**	La experiencia del experto está poco organizada.	D
TA	A2	6	**10**	Tiene valor práctico.	D
TA	A3	7	**10**	Es una tare más práctica que estratégica.	D
TA	A4	7	**10**	La tarea da soluciones que sirvan a necesidades a largo plazo.	E
TA	A5	5	**10**	La tarea no es demasiado fácil, pero es de conocimiento intensivo, tanto propio del dominio, como de manipulación de información.	D
TA	A6	6	**10**	Es de tamaño manejable, y/o es posible un enfoque gradual y/o, una descomposición en subtareas independientes.	D
TA	A7	7	**10**	La transferencia de experiencia entre humanos es factible (experto a aprendiz).	E
TA	A8	6	**9**	Estaba identificada como un problema en el área y los efectos de la introducción de un SE pueden planificarse.	D
TA	A9	9	**7**	No requiere respuestas en tiempo real inmediato.	E
TA	A10	9	**7**	La tarea no requiere investigación básica.	E
TA	A11	5	**10**	El experto usa básicamente razonamiento simbólico que implica factores subjetivos.	D
TA	A12	5	**10**	Es esencialmente de tipo heurístico.	D

Tabla 3.5: Éxito.

CAT.	IDEN. CAT.	PESO (P)	VALOR (V)	DENOMINACIÓN DE LA CARACTERÍSTICA	TIPO
EX	E1	8	**10**	No se sienten amenazados por el proyecto, son capaces de sentirse intelectualmente unidos al proyecto.	D
EX	E2	6	**10**	Tienen un brillante historial en la realización de esta tarea.	D
EX	E3	5	**10**	Hay acuerdos en lo que constituye una buena solución a la tarea.	D
EX	E4	5	**10**	La única justificación para dar un paso en la solución es la calidad de la solución final.	D
EX	E5	6	**7**	No hay un plazo de finalización estricto, ni ningún otro proyecto depende de esta tarea.	D
TA	E6	7	**7**	No está influenciada por vaivenes políticos.	E
TA	E7	8	**6**	Existen ya SSEE que resuelvan esa o parecidas tareas.	D
TA	E8	8	**5**	Hay cambios mínimos en los procedimientos habituales.	D
TA	E9	5	**10**	Las soluciones son explicables o interactivas.	D
TA	E10	7	**10**	La tarea es de I+D o de carácter práctico, pero no ambas cosas simultáneamente.	E
DU	E11	6	**10**	Están mentalizados y tienen expectativas realistas tanto en el alcance como en las limitaciones.	D
DU	E12	7	**10**	No rechazan de plano esta tecnología.	E
DU	E13	6	**10**	El sistema interactúa inteligente y amistosamente con el usuario.	D
DU	E14	9	**10**	El sistema es capaz de explicar al usuario su razonamiento.	D
DU	E15	8	**10**	La inserción del sistema se efectúa sin traumas; es decir, apenas se interfiere en la rutina cotidiana de la empresa.	D
DU	E16	6	**10**	Están comprometidos durante toda la duración del proyecto, incluso después de su implantación.	D
DU	E17	8	**10**	Se efectúa una adecuada transferencia tecnológica.	E

CAPÍTULO 4

ESTADO DEL ARTE

Este capítulo pretende proporcionar una visión global de la investigación bibliográfica llevada a cabo para comprender el proceso de desarrollo del prototipo computacional para la detección de mensajes de odio en la red.

Este trabajo ha requerido que dos grandes áreas tan dispares como el Derecho Penal y la Inteligencia Artificial se entiendan para dar como fruto un sistema que utilizando técnicas de Análisis de Sentimientos y Lógica Borrosa sea capaz de discernir si existe odio en un mensaje y su intensidad.

Haciendo uso de estas dos grandes áreas de conocimiento se presenta un estado del arte centrado en el tema a abordar. Por parte del derecho penal se expone una visión histórica del delito de odio, centrándonos en la jurisprudencia del mismo en el estado Español, sin olvidar el objetivo de análisis, que son mensajes de odio en la red y las medidas que se adoptan para proteger al ciudadano de esta actividad delictiva, lo que conlleva al uso de técnicas de IA, en concreto el Procesamiento del Lenguaje Natural y en Sistemas de Basados en el Conocimiento que permiten la clasificación y detección de los mismos. Por último se dará una visión global de la Lógica Borrosa para presentar todo el conocimiento sobre el que se sostiene el proyecto (Ver Figura 4.1).

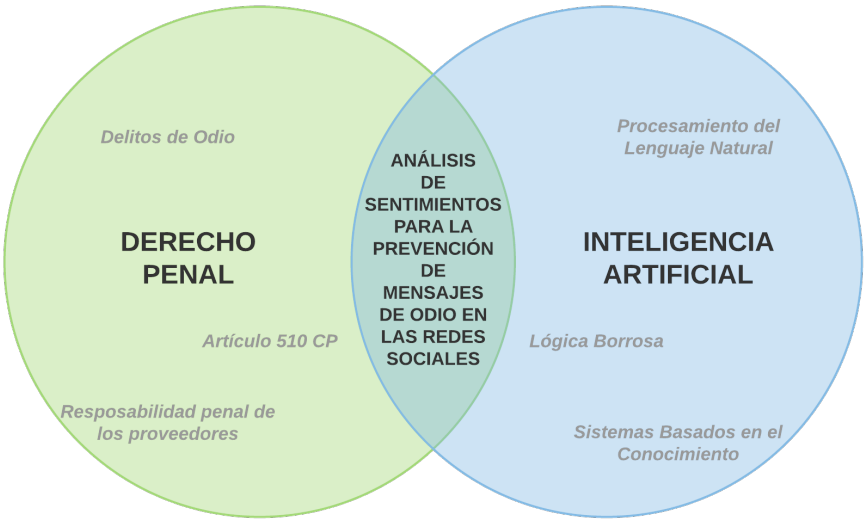


Figura 4.1: Áreas de conocimiento directamente relacionadas con el proyecto.

4.1 DERECHO PENAL: LOS DELITOS DE ODIO

Para entender el Delito de Odio propiamente dicho, primero hemos de explorar su matriz terminológica partiendo de la Convención Europea de Derechos Humanos[25] y de la historia legislativa Estadounidense.

4.1.1 Historia: Matriz terminológica del Discurso del Odio

Europa

No existe una definición universalmente aceptada de "Discurso de Odio" según Weber en [78], aunque la mayoría de los estados que componen Europa han adoptado medidas que prohíben de una forma u otra el uso de expresiones de odio, los preceptos difieren ligeramente en lo que realmente prohíben. Weber apunta que la definición más genérica de "*Discurso de Odio*", es la Recomendación (97) 20 del Comité de Ministros del Consejo de Europa[50] aprobado el 30 de octubre de 1997. La definición reza así: «*El término Discurso de Odio se entenderá que abarca toda clase de expresión que difunde, incita, promueve o justifica el odio racial, la xenofobia, el antisemitismo u otras formas de odio basadas en la intolerancia incluida la intolerancia que se manifiesta a través del nacionalismo agresivo y el etnocentrismo, la discriminación y la hostilidad contra las minorías, migrantes y personas de origen inmigrante*».

Esta definición de referencia a en el marco Europeo se ha actualizado para incluir a colectivos minoritarios no amparados en la anterior definición. Esta nueva definición de referencia se acuñó en la Comisión Europea contra el Racismo y la Intolerancia (ECRI) el 8 de diciembre de 2015 [60]. La actualización del concepto se presenta de la siguiente forma: «*el discurso de odio debe entenderse como fomento, promoción o instigación, en cualquiera de sus formas, del odio, la humillación o el menosprecio de una persona o grupo de personas, así como el acoso, descrédito, difusión de estereotipos negativos, estigmatización o amenaza con respecto a dicha persona o grupo de personas y a la justificación de esas manifestaciones por razones de "raza", color, ascendencia, origen nacional o étnico, edad, discapacidad, lengua, religión o creencias, sexo, género, identidad de género, orientación sexual y otras características o condiciones personales;*».

Por tanto el Delito de Odio debe entenderse como la criminalización del Discurso del Odio que cada estado europeo castiga en sus respectivos Códigos Penales.

Estados Unidos

A diferencia de Europa en la que la legislación contra el Delito de Odio tuvo como objetivo inicial la "*desnazificación*", en Estados Unidos, este conato de legislación se produjo después de la Guerra de Secesión en la que el concepto de discriminación fue garantía de protección en la Decimocuarta Enmienda de 1868[1]. A partir de ese momento aparecieron nuevos estatutos como el "*The Ku Klux klan Act*" de 1871[2] en el que se castigaba penalmente a funcionarios del gobierno y a conspiradores tanto civiles como privados que cometían actos discriminatorios contra esclavos recién emancipados y a otros grupos desiguales y

[1] http://www.loc.gov/rr//program/bib/ourdocs/14thamendment.html
[2] http://legisworks.org/sal/17/stats/STATUTE-17-Pg13.pdf

posteriormente el conocido como la acta de derechos civiles ("*Civil Rights Acts*") de 1875[3] que garantizaba que cualquier ciudadano independientemente de su raza, color o previa condición de esclavitud, de poder disfrutar de forma plena y equitativa los alojamientos públicos, medios de transporte y lugares de ocio [42].

A pesar de todo en 1883 el Tribunal Supremo invalidó "*The Civil Rights Acts*" en 1883, como consecuencia todas las enmiendas posteriores a la Guerra de Secesión resultaron inútiles y no fue hasta 1966 que la Corte Suprema de Estados Unidos aprobó la aplicación de leyes penales a favor de los derechos civiles[4,5]. A partir de ese momento fue el preludio para la legislación de leyes anti-odio.

En 1990, ya a nivel federal se crea la Ley Estadística de Delitos de Odio ("*Hate Crime Statistics Act*")[6] aprobada por el Congreso estadounidense con el objetivo de recolectar y publicar datos sobre crímenes que manifiesten claramente objeto de odio basado en la raza, la religión, la etnia y la orientación sexual, cuyo mandato legal acabó asumiendo el FBI[58] que se incluyó en el *Uniform Crime Reporting Program* (UCR[7]). Pero el cambio más significativo y que supuso la universalización de los Delitos de Odio, fue La Ley de Delitos de Odio de 2009 denominada Matthew Shepard y James Byrd[8] que asienta una amplia competencia federal en la materia. Esta ley tipifica como delito el causar intencionadamente lesiones corporales o el intento de las mismas, motivadas por la raza, color, religión, etnia, sexo, orientación sexual, identidad de género o discapacidad real o percibida de cualquier persona, siendo la primera ley que permite el enjuiciamiento penal a nivel federal por delitos de odio motivados por su orientación sexual o identidad de género real o percibida.

Se observa una clara diferencia entre la matriz terminológica europea y americana y su posterior aplicación penal, ya que en Europa se establece lo que se conoce como Discurso de Odio (*"hate speech"*) y da la competencia a los estados para legislar sobre el delito, en cambio en Estados Unidos se habla de Delito de odio (*"hate crime"*) porque conlleva un crimen o intento del mismo por las motivaciones antes nombradas.

España

La matriz terminológica de la expresión "Delitos de Odio" en el Estado español, viene dada por la interpretación establecida por el Tribunal Europeo de Derechos Humanos respecto al Discurso de Odio (*"hate speech"*) y a la evolución político-criminal del Delito de Odio (*"hate crime"*) en los Estados Unidos.

En España todas las referencias terminológicas apuntan a la protección de determinados colectivos o minorías con un determinado grado de vulnerabilidad, es decir, los referentes terminológicos aluden que los ataques penalmente relevantes son a grupos que históricamente arrastran cierto estigma en términos de marginación, vulnerabilidad, hostilidad, o cualquier forma de discriminación por razón de pertenencia a ese grupo social, lo que incluye la protección a grupos sociales mayoritarios como las mujeres, aunque inicialmente la matriz terminológica apunte a colectivos minoritarios [38].

[3] http://legisworks.org/sal/18/stats/STATUTE-18-Pg335a.pdf
[4] https://supreme.justia.com/cases/federal/us/383/745/
[5] https://supreme.justia.com/cases/federal/us/383/787/
[6] https://ucr.fbi.gov/hate-crime/2011/resources/hate-crime-statistics-act
[7] https://www.fbi.gov/services/cjis/ucr
[8] https://www.justice.gov/crt/matthew-shepard-and-james-byrd-jr-hate-crimes-prevention-act-2009-0

Esta designación que apunta a los delitos cometidos con palabras o hechos confluyen tanto en España como a nivel europeo (por razón de pertenencia a la comunidad europea) en la **Decisión Marco 2008/913/JAI del Consejo, de 28 de noviembre de 2008 relativa a la lucha contra determinadas formas de manifestaciones de racismo y xenofobia mediante el derecho penal**[9], que en síntesis garantiza que determinadas manifestaciones graves de racismo y xenofobia sean castigadas con sanciones penalmente efectivas en toda la Unión Europea (UE), lo que no incluye la protección penalmente relevante por discriminación (puede o no implicar incitación a la violencia) a colectivos por género, orientación o identidad sexual, enfermedad o discapacidad, que si son recogidos como normativa penal complementaria en el Código Penal (CP) del Estado español en los artículos 510, 510 bis, 515.4º y 22.4º.

4.1.2 España: Discurso de Odio Criminalizado

Con la entrada en vigor del actual CP en 1995, la normativa penal anti-odio experimentó un cambio sin precedentes con una marcada vocación de tutela antidiscriminatoria y de principio de igualdad, donde se incorporaron figuras penales como el agravante genérico que castiga la comisión de cualquier delito propiciado por motivos racistas, étnicos, ideológicos, religiosos, sexo, condición sexual, género, enfermedad o discapacidad como es el artículo 22,4º (Ver Anexo A). Hasta un precepto que castiga la incitación o fomento del odio por las razones antes expuestas como es el artículo 510 del CP (Ver Anexo A).

Con la última reforma del CP llevada a cabo por la Ley Orgánica 1/2015, de 30 de marzo, se modifica el artículo 510 del CP para entre otros motivos adaptarse a la decisión marco de 2008 de la UE y se añade el artículo 510 bis del CP.

Es bien conocido que no solo el artículo 510 del CP se ocupa de las prohibiciones penales del discurso de Odio, también existen preceptos que pueden ser motivados por el odio como por ejemplo los relativos a la violencia de género (art. 153.2, 173.2 ... CP), delitos de manipulación genética para la creación de seres humanos idénticos por motivos étnicos (art. 160.3 CP), los crímenes contra la humanidad (art. 607 bis CP), etc. Pero como se ha nombrado anteriormente es el artículo 510 del CP el que hace referencia en su sentido más estricto a las prohibiciones penales del discurso de odio.

Artículo 510

Tras la última reforma del CP en 2015 se pueden observar hasta 6 figuras penales diferentes que componen el art. 510 del Código Penal (Ver Anexo A). Pero en síntesis se pueden distinguir dos bloques claramente diferenciados que ya se distinguían en el CP del año 95. Un primer bloque que castiga la incitación pública grave de comunicación de odio, entendiéndose como la antesala al acto por dicho motivo que sería castigado en el agravante 22.4 del CP. Y el segundo bloque de tipos atenuados con respecto al primero, que castiga las conductas injuriosas que lesionen la dignidad de las personas o enaltezcan cualquier medio de expresión pública o de difusión por motivos de Odio. Estos bloques hacen referencia a los artículos 510.1 y 510.2 respectivamente. El resto de preceptos del artículo 510 pueden entenderse como una serie de agravantes o matices terminológicas a los

[9] https://goo.gl/S67q3W

dos primeros preceptos. A continuación se muestra un análisis de las figuras penales antes mencionadas:

- **Matriz de incitación**: Los preceptos que componen este bloque de conductas son considerados incitación al odio grave ya que pueden ser castigados con penas de prisión de uno a cuatro años y multa de seis a doce meses.
  - **Artículo 510.1.a**: Esta primera modalidad puede denominarse delito de incitación pública al odio (Ver Anexo A). Fomentar, promover o incitar son verbos que en términos jurídico-penales implican a terceras personas en el acto de odiar o discriminar contra miembros del colectivo diana (colectivos que protege el artículo 510 del CP). Como apunta Landa en [39] *«(…) La relevancia penal de la conducta se alcanza, sin embargo, cuando el contenido tendencial es de tal intensidad que puede colegirse con claridad que la hostilidad, el odio, la violencia o la discriminación se despliegan como medios eficaces para promover, fomentar o incitar su repetición a una escala que pueda llegar a afectar el ejercicio de derechos fundamentales (…) de los miembros del colectivo contra el que el discurso se despliega. (…)»*. Por tanto se llega a la conclusión de que la gravedad de la pena que implica este precepto se contempla en aquellos discursos que se propaguen de forma pública o por medios masivos (ver precepto 510.3 en Anexo A), excluyendo aquellos que se produzcan en esferas privadas o semi-privadas.
  - **Artículo 510.1.b**: La miscelánea de casos que tienen relevancia en este precepto hacen alusión a la producción y elaboración de "productos" de odio o con objeto de llevar a cabo estas acciones para su posterior distribución en medios de difusión (ver Anexo A). Este artículo solo alcanzará la relevancia establecida en la pena del precepto, cuando el contenido se pueda identifica de manera inteligible una clara intención de odio, discriminación e incitación a la violencia que pueda afectar al ejercicio de los derechos fundamentales del colectivo diana [39].
  - **Artículo 510.1.c**: Este artículo (ver Anexo A) nace de la reforma del CP de 2015, tomando parte de la tipificación de las conductas que pertenecían al art. 607.2. Con este precepto se pretende adelantar la intervención penal de condiciones sociales que favorezcan (mediante la palabra o el escrito) la creación de discursos que se articulan en torno a conductas de apologías de crímenes de derecho penal internacional [18].
- **Matriz injuriosa**: La modalidad de la que compete el art. 510.2 del CP parten de conductas injuriosas con similitud a lo descrito sobre el delito de incitación grave del art. 510.1.a del CP, a su cadena de difusión en el art. 510.1.b del CP y a la apología de crímenes de derecho internacional en el art. 510.1.c, con la diferencia que estas prohibiciones parten de un castigo más leve con penas de prisión de seis meses a dos años y multa de seis a doce meses.
  - **Artículo 510.2.a**: Se protege la afrenta contra el colectivo diana (Ver Anexo A) siempre y cuando sea de tal intensidad que entrañe un gran potencial agresivo, ponga en peligro el ejercicio de sus libertades y atente contra la dignidad humana dicho colectivo [39].
  - **Artículo 510.2.b**: Este precepto prohíbe conductas apologéticas por cualquier medio de difusión publica de delitos. Como analiza el profesor Landa en [39],

existe una gran similitud con el art. 510.1.c del CP sobre conductas que hacen apología de delitos de derecho penal internacional, por tanto la diferencia entre ambos preceptos es, la pena que se interpone en cada uno, siendo más leve la del artículo actualmente analizado, en síntesis, se impondrá una pena u otra según la intensidad del delito intentando contextualizar la conducta según la fuerza incitadora.

- **Modalidades complementarias**: Preceptos que interfieren en la pena de forma agravada o precisando la condena de los artículos 510.1 y 510.2 del CP.

  - **Artículo 510.3**: Esta modalidad hace referencia a el medio en el que se comete el Delito de Odio, en concreto a los delitos cometidos "«*a través de un medio de comunicación social, por medio de internet o mediante el uso de tecnologías de la información*»" (Ver Anexo A). Como consecuencia del uso de estos medios de difusión y asumiendo que se demuestra el alcance masivo por el motivo inicial, la pena se agrava a su mitad superior prevista en los artículos 510.1 y 510.2 del CP.
  - **Artículo 510.4**: El tipo cualificado de este artículo hace referencia a un agravante de pena de los sub-tipos 510.1 y 510.2 elevándola a la mitad superior con posibilidad de incremento a la máxima establecida en el tipo básico 510.1 del CP, a saber, hasta 6 años de prisión. Para la aplicación de este artículo se debe percibir una alteración de la paz pública, la creación de un grave sentimiento de inseguridad y temor por la integridad del colectivo diana. Es por tanto una situación próxima al conflicto colectivo [39].
  - **Artículo 510.5**: Determina de forma preceptiva la pena inhabilitación docente de entre tres y diez años añadido al de la pena de privación de libertad, en el caso que exista, a los profesionales educativos y deportivos y de tiempo libre con el objeto de protección de los menores como potenciales destinatarios de propaganda de odio [38].
  - **Artículo 510.6**: Este artículo habilita al juez a la destrucción, a cualquier clase de soporte de objeto del delito, incluyendo a supuestos en que se hubiera cometido el delito a través de medios y tecnologías de la información.

4.1.3 Diagnóstico y actuación frente a los Delitos de Odio

La detección de infracciones cometidas por el odio o aversión irracional hacia un colectivo diana son complejas de investigar, ya que un precepto tan complejo requiere de un conocimiento por parte de todo el proceso de investigación (por parte de las fuerzas de seguridad del estado, e incluso el servicio de emergencias del 112) y enjuiciamiento (los fiscales y jueces que deben instruir la causa, hasta los jueces y magistrados que deben valorar los hechos y dictar sentencia). Otra dificultad añadida es debido a la reciente reforma del art. 510 del CP que prácticamente no ha tenido tiempo de ser aplicada por los tribunales lo que conlleva que no se haya producido un cuerpo de jurisprudencia suficientemente depurado [1]. Y por último, dictar una sentencia por Delito de Odio es altamente complejo ya que es fácil que entre en conflicto con el derecho fundamental de la Libertad de Expresión (art. 10.1 CEDH Ver Anexo D) y por tanto se aplique el art. 17 del CEDH (Ver Anexo D) quedando exento del supuesto delito.

Identificación de los Delitos de Odio

El objetivo de los Delitos de Odio es destruir la convivencia mediante una animadversión irracional hacia colectivos vulnerables. Las casuísticas posibles a la hora de identificar esta tipología de delito pueden ser diversas y descubrir la motivación del acto es condición indispensable [1].

Por tanto, se hace imprescindible definir una serie de indicadores con el objetivo de dotar a los jueces y fiscales de información suficiente para permitirles dictar sentencia. La sentencia del TEDH de 20/10/2015 Balázs versus Hungría[10] avala la validez de los indicadores que se enumeran a continuación:

1. La percepción de la víctima.
2. La pertenencia de la víctima a un colectivo o grupo minoritario.
3. Discriminación y odio por asociación.
4. Las expresiones o comentarios racistas, xenófobos u homófobos, o cualquier otro comentario vejatorio contra cualquier persona o colectivo, por su ideología, situación de exclusión social, orientación religiosa, por ser persona con discapacidad, etc.
5. El vestuario, tatuajes o la estética del autor de los hechos.
6. La propaganda, estandartes, banderas, pancartas, etc. de carácter extremista o radical.
7. Los antecedentes policiales del sospechoso.
8. La localización del incidente.
9. La relación del sospechoso con los grupos ultras de fútbol.
10. La asociación del sospechoso con asociaciones que favorezcan el odio.
11. La aparente gratuidad de los actos violentos.
12. Enemistad histórica entre los miembros del grupo de la víctima y del presunto culpable.
13. Cuando los hechos ocurran en un día, hora o lugar en el que se conmemora un acontecimiento o constituye un símbolo para el delincuente.
14. Si los hechos ocurren con motivo u ocasión de una fecha significativa para la comunidad o colectivo de destino.
15. La conducta del infractor.

Este manual de indicadores de polarización más frecuente en Delitos de Odio forma parte del **Protocolo de actuación de las Fuerzas y Cuerpos de Seguridad del Estado para los Delitos de Odio y conductas que vulneran las normas legales sobre discriminación**[11] y del **Manual práctico para la investigación y enjuiciamiento de Delitos de Odio y Discriminación**[1], para una información más precisa ver Anexo E.

[10] http://hudoc.echr.coe.int/fre/?i=001-158033

[11] http://www.interior.gob.es/documents/642012/3479677/PROTOCOLO+ACTUACION/99ef64e5-e062-4634-8e58-503a3039761b

Evaluación de los Delitos de Odio

La interpretación del art. 510 del Código Penal, en su nuevo tenor literal, castiga la incitación directa e indirecta, por tanto, como se vio en la sección §4.1.2 no solo se puede colegir del contenido del discurso, ni de la crudeza de las palabras empleadas el delito, tiene que producirse en un contexto en el que sea susceptible la agitación colectiva para establecer un clima del paso al acto, es decir, el Discurso de Odio tiene que tener un alcance público, porque como se señala en [54] la incitación o provocación que se cometa de forma privada o semi-privada sin capacidad extensiva seguirá siendo impune.

Para determinar la severidad del Discurso del Odio y establecer una relevancia jurídico-penal sólida para la toma de decisiones penales frente al delito se propone el test de relevancia denominado **Plan de Acción de Rabat** (ver Anexo F) que proporciona los criterios para contextualizar el Discurso de Odio y establecer el nivel de gravedad de las mismas. El test de severidad ha sido sintetizado por la **Recomendación de Política General número 15 relativa a la lucha contra el Discurso de Odio**[28] en el que se evalúa el riesgo del paso al acto en los siguientes puntos:

(a) *"el contexto en el que se utiliza el discurso de odio en cuestión (especialmente si ya existen tensiones graves relacionadas con este discurso en la sociedad)"*;

(b) *"la capacidad que tiene la persona que emplea el discurso de odio para ejercer influencia sobre los demás (con motivo de ser por ejemplo un líder político, religioso o de una comunidad)"*;

(c) *"la naturaleza y contundencia del lenguaje empleado (si es provocativo y directo, si utiliza información engañosa, difusión de estereotipos negativos y estigmatización, o si es capaz por otros medios de incitar a la comisión de actos de violencia, intimidación, hostilidad o discriminación)"*;

(d) *"el contexto de los comentarios específicos (si son un hecho aislado o reiterado, o si se puede considerar que se equilibra con otras expresiones pronunciadas por la misma persona o por otras, especialmente durante el debate)"*;

(e) *"el medio utilizado (si puede o no provocar una respuesta inmediata de la audiencia como en un acto público en directo)"*;

(f) *"la naturaleza de la audiencia (si tiene o no los medios para o si es propensa o susceptible de mezclarse en actos de violencia, intimidación, hostilidad o discriminación)"*.

4.1.4 Comunicación Violenta y de Odio en las Redes Sociales

Como se expuso en el capítulo §1 Internet ha supuesto un nuevo universo para la expansión del odio incrementando la facilidad de acceder a dicho contenido a miles de usuarios.

La aparición de las Redes Sociales ha supuesto un cambio en el paradigma de entendimiento de las comunicaciones, lo que ha permitido que las comunicaciones interpersonales a través de la red alcancen a millones de personas en un corto periodo de tiempo [15]. Este nuevo paradigma de comunicación ha provocado la expansión del discurso de odio

a través de las Redes Sociales potenciando así sus efectos comunicativos y su capacidad incitación en la comunidad [27], lo que puede conllevar también a hacer uso de estas herramientas de comunicación un foro de radicalización violenta usado por grupos terroristas para reclutamiento o difusión de mensajes de odio [72][16] [67][74].

El potencial transformador de las Redes Sociales plantea múltiples desafíos a nivel internacional, de los cuales la incitación al odio ha sido reconocida por múltiples estados como un gran problema social. Fruto de esta preocupación nace *No Hate Speech Movement*[12] que es una iniciativa del *Council of Europe of Youth Department* que pretende movilizar a los jóvenes para combatir el odio y promover los derechos humanos en Internet. También la UNESCO publicó un estudio *Countering online hate speech* [30] con el objetivo de proporcional un manual de buenas prácticas para ayudar a los países a hacer frente al problema del odio en la red.

En España esta preocupación social internacional por el discurso de odio en Internet llevó, entre otros motivos, a la Fiscalía General del Estado a la última reforma del CP llevada a cabo por la Ley Orgánica 1/2015, de 30 de marzo [64].

Las Redes Sociales parecen el medio idóneo para la propagación del discurso del odio por su capacidad de llegar a una gran multitud de personas y por el potencial anonimato que proporciona lo que puede conllevar a una ilusión de impunidad frente a la ley.

4.2 INTELIGENCIA ARTIFICIAL

Como se ha expuesto en las secciones anteriores la necesidad de detectar CVydO en la red se ha vuelto crucial, ya que este fenómeno de odio online tiene el poder de incitación capaz de crear un sentimiento de odio generalizado contra los diferentes colectivos diana poniendo en peligro su integridad y el ejercicio de sus derechos.

Para ello las Tecnologías de la Información (TI) pueden ser útiles herramientas en las distintas fases de detección y prevención del odio en la red, desde lanzar campañas preventivas a través de medios sociales hasta la detección del Discurso del Odio mediante técnicas de Inteligencia Artificial (IA en adelante).

Este capítulo pretende mostrar una visión general del papel que desempeña la IA en la detección del Discurso de Odio en la red.

4.2.1 Inteligencia Artificial: una visión general

El concepto actualmente conocido como Inteligencia Artificial (IA) nace en el año 1956 propuesto por John McCarthy y aprobado en el congreso de Dartmouth [46].

No existe una definición única válida de IA. Peter Norvig y Stuart Rusell en [65] hacen una clasificación de cuatro enfoques proporcionando varias definiciones extraídas de distintas fuentes (Ver Tabla 4.1). A continuación se exponen brevemente cada uno de los enfoques:

- **Pensando como humanos**: Este enfoque pretende enfocar la IA como modelos cognitivos [80], es decir, imitar en una máquina cómo piensan los humanos. Pero

[12] https://www.coe.int/en/web/no-hate-campaign

Tabla 4.1: Definiciones de Inteligencia Artificial recopiladas en [65].

Pensando como humanos	**Pensando Racionalmente**
«"*El nuevo y excitante esfuerzo de hacer que los computadores piensen (…) máquinas con mentes, en el más amplio sentido literal.*"»[33] «"*La automatización de actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje (…).*"»[3]	«"*El estudio de las facultades mentales mediante el uso de modelos computacionales.*"»[29] «"*El estudio de los cálculos que hacen posible percibir, razonar y actuar.*"»[81]
Actuando como humanos	**Actuando racionalmente**
«"*El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia.*"»[37] «"*El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor.*"»[63]	«"*La Inteligencia Computacional es el estudio del diseño de agentes inteligentes.*"»[59] «"*IA (…) está relacionada con conductas inteligentes en artefactos.*»" [56]

antes se debe conocer como trabaja la mente mediante experimentos psicológicos o mediante introspección.

- **Pensando racionalmente**: La IA como mecanismo computacional de resolución de problemas usando la lógica, es decir, formalizando el conocimiento y expresándolo en notación lógica.

- **Actuando como humanos**: El padre de este enfoque es **Alan Turing**, el cual diseñó lo que se conoce como el **Test de Turing** (1950) cuyo objetivo para proporcionar la definición de inteligencia era la incapacidad de diferenciar entre una máquina y un ser humano. La características que debía tener una máquina para superar la prueba son consideradas las disciplinas que abarcan la mayor parte de la IA actual. A saber:

  - **Procesamiento del lenguaje natural.**
  - **Aprendizaje automático.**
  - **Representación del conocimiento.**
  - **Razonamiento automático.**
  - **Visión por computador.**
  - **Robótica.**

- **Actuando racionalmente**: Este enfoque compete al mundo de los agentes racionales, que son aquellos que actúan de forma autónoma con la intención de alcanzar una meta y la capacidad de tolerancia a fallos lo que les confiere la capacidad de gestionar la incertidumbre y proporcionar el mejor resultado posible a un problema dado.

4.2.2 Procesamiento del Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN en adelante) es un área de investigación y aplicación de la IA, la cual tiene como objetivo desarrollar sistemas computacionales capaces de manipular y entender el texto o el habla en lenguaje natural.

El PNL engloba múltiples disciplinas de aplicación tales como la traducción automática de texto, síntesis automática de texto, análisis de sentimientos, dictado por voz a texto y viceversa, etc.

Para la detección de CVydO en la red el PNL proporciona mecanismos útiles para la identificación de comportamientos de odio en las publicaciones en red. A continuación se muestra una revisión de los estudios que hacen uso de esta disciplina para detectar mensajes de odio.

Enfoque basado en las características del mensaje

- **Bolsas de términos:**

  La forma más básica de clasificar un mensaje es atendiendo a las palabras que contiene, es decir, con el uso de bolsas de términos, la gran mayoría de investigadores hacen uso desde unigramas hasta n-gramas para la identificación de mensajes de odio. En [20] muestra un experimento en que se comparan seis enfoques para la precisión en la predicción de frases ofensivas de los cuales cuatro están directamente relacionados con las bolsas de términos. Otras obras como [83] [70] [10] [75] [77] [11] [57] hacen uso de bolsas de términos compuestas por n-gramas para la detección de mensajes de odio, enriqueciendo sus modelos con características adicionales.

  Este enfoque aunque efectivo, plantea una serie de dificultades, ya que cualquier variación ortográfica podría plantear un problema por no contemplarse en la bolsa de términos aun aplicando algoritmos de "*stemming*" para reducir las palabras a la raíz, o técnicas para la corrección ortográfica como en [20] que se ayudan del corpus de WordNet[13] y de un algoritmo de corrección ortográfica[14] para homogeneizar sus datos.

- **Generalización:**

  Otro problema es la carencia de datos para cubrir el espacio de entrada y salida de los modelos. Esto viene propiciado por la naturaleza del entorno donde se dan estos fenómenos de odio en la red, a saber, como se ha introducido en la sección anterior (§4.1.4), la aparición de las Redes Sociales han universalizado el discurso del odio, precisando, con la aparición del "*microblogging*" se ha abierto un nuevo paradigma de comunicación, que consiste en un breve texto en el que los usuarios de ese servicio pueden describir su estado actual en publicaciones distribuidas en la Web [35], como por ejemplo la Red Social Twitter en la que el número máximo de caracteres permitidos actualmente es de 280 lo que en ocasiones impide la multiplicidad terminológica dentro del mensaje que están más orientados a ser concisos en lo que al contenido se refiere.

  Por tanto, para la escasez de términos algunas obras abordan este problema recurriendo a la generalización de palabras, que consiste en agrupar términos en conjuntos

[13] https://wordnet.princeton.edu
[14] http://norvig.com/spell-correct.html

representativos de los mismos. En [76] se hace uso del algoritmo propuesto por [8] en el que se asigna términos a agrupaciones de los mismos. Existen otras formas de generalizar haciendo uso del algoritmo Latent Dirichlet Allocation (LDA) [4] donde la agrupación de las palabras (o n-gramas) se hace por grado de pertenencia, por tanto cada palabra o grupo de palabras puede pertenecer a mas de una clase (o cluster), en [82] hacen uso de dicho algoritmo para agrupar en temas una serie de tweets con el propósito de detectar contenido ofensivo en Twitter concluyendo que su modelo obtiene mejores resultados que uno basado en la identificación únicamente de palabras clave.

Otros enfoques para la generalización de términos están basados en el uso de Redes Neuronales para la representación de modelos vectoriales de palabras distribuidas [49] usado en [57]. En [27] hacen uso de un marco no supervisado que aprende de representaciones vectoriales distribuidas para fragmentos de texto de longitud variable [41] con el objetivo de obtener a partir de un texto palabras semánticamente similares donde demuestran una mejor precisión que enfoques basados meramente en bolsas de términos.

- **Aspectos lingüísticos:**

  Las características lingüísticas que se pueden extraer del mensaje desempeñan una labor significativa en la detección de incitación al odio.

  En el sistema *Smokey* de [71] concibe una serie de reglas basadas en la sintaxis y en la semántica de cada mensaje. Las características sintácticas incluyen la detección de frases nominales como aposiciones o la detección de imperativos, también incorpora reglas semánticas para evitar los falsos positivos incorporando una serie de "*reglas de elogios*", con el fin de evitar clasificar frases que contengan esta serie de palabras (que forman las "*reglas de elogios*") como negativas.

  En [83] enriquece su modelo de n-gramas con un etiquetado gramatical (POS) haciendo uso del software de PLN Stanford CoreNLP [15]. En [20] hace un análisis de la dependencia sintáctica dentro de la oración identificando a quién va dirigida la ofensa construyendo así una serie de reglas para la detección de lenguaje ofensivo. También hacen uso del análisis de dependencias sintácticas [32] [10] [11] y [57]. Mientras que [20] y [32] seleccionan una de las características del análisis de dependencia de "forma manual". En [10] aplican una selección de características mediante una Regresión Logística Bayesiana (BLR) para determinar qué características eran más significativas estadísticamente para clasificar el odio en la red.

Uso de recursos para la identificación de mensajes de odio

Es bien conocido que para la identificación del discurso del odio es necesario detectar palabras ofensivas tales como insultos, amenazas, calumnias, etc. Para ello existen recursos web que son empleados por varios autores como [82] [10] y [57] que usan corpus compuestos por términos considerados generales relacionados con el odio[16,17,18] respectivamente. Como contraposición al uso de corpus generalistas para la detección de términos de odio, [11]

[15] https://stanfordnlp.github.io/CoreNLP/
[16] www.noswearing.com/dictionary
[17] www.rsdb.org
[18] www.hatebase.org

se centra en diccionarios de palabras especializadas en subtipos particulares de odio como injurias contra el origen étnico[19], por razón de pertenencia al colectivo LGTBI[20] o por el mero hecho de sufrir una discapacidad[21].

Otros autores emplean listas de términos específicamente creadas para sus trabajos como es el caso de [71] con sus listas de expresiones regulares compuestas por los llamados "bad-verbs" ("stink", "suck", etc), "bad-nouns"("loser", "idiot", etc), etc. En [62] crean un diccionario con alrededor de 2.700 palabras, frases y expresiones al que llaman "*Isulting or Abusive Language Dictionary*", donde clasifican de forma manual en una escala de 1 a 5 según su nivel de odio y posteriormente ajustan esas ponderaciones mediante aprendizaje adaptativo. En [32] extraen a partir de 6 verbos incitatorios principales una recopilación todos los verbos de odio que se encuentran en el corpus que utilizan, para ello utilizan un algoritmo de *bootstrapping*, la funcionalidad "synsets" de WordNet (que proporciona a partir de una palabra una serie de palabras semánticamente similares según el contexto) e hiperónimos para construir esta lista de verbos incitatorios.

Uso de meta-información para la detección del discurso del odio

La meta-información o información sobre información es un recurso valioso para la detección de comentarios de odio en la red ya que el texto a analizar proviene casi exclusivamente de RRSS y gracias a sus APIs se puede obtener meta-información útil para el análisis de este tipo de mensajes.

Como se especifica en el capítulo §6 sección §6.3 la persona reincidente en comentarios injuriosos y de incitación a la violencia, se puede afirmar que es más propenso a cometerlos de nuevo, por tanto, conocer el comportamiento de un usuario específico dentro de una red social puede ser meta-información útil para la detección de mensajes de odio. En [82] hace uso de esta heurística, mientras que [24] utiliza como metadatos el número de palabras de odio que hay en el historial de mensajes de un usuario. En [23] y [77] concluyen que los hombres son más propensos a cometer actos de odio.

A parte de la meta-información intrínseca al mensaje, se puede extraer otro tipo de metadatos a partir de las APIs de las RRSS. En [53] utiliza meta-información extraída de los usuarios de Twitter como por ejemplo si la cuenta está verificada o no, si incluye biografía, el número de días desde la creación de la cuenta, el número de seguidores, el número de seguidos, etc; y de los tweets que estos publican como el número de menciones, si hacen uso de *"hashtag"*, el número de *"retweets"*, etc; de un corpus de tweets etiquetados extraídos a raíz del atentato de Charlie Hebdo en [52] para clasificar los cometarios violentos y de odio de los que no lo son. En [77] hace uso del origen geográfico de los usuarios como meta-información.

Métodos de clasificación para la detección del discurso del odio

El enfoque predominante para la detección de discursos de odio es el de aprendizaje supervisado. La mayoría de autores hacen uso del clasificador "*Support Vector Machine" (SVM)*[22], a saber, [82] [23] [24] [12] [13] [76] [75] [83] con el fin de identificar mensajes de

[19] https://en.wikipedia.org/wiki/List$_of_ethnic_slurs$

[20] https://en.wikipedia.org/wiki/List$_of_LGBT_slang_terms$

[21] https://en.wikipedia.org/wiki/List$_of_disability-related_terms_with_negative_connotations$

odio de los que no lo son. En [53] hacen uso de *Random Forest classifier* [7] incluyendo las características del usuario y del mensaje para clasificar los tweets en Neutros o en Discurso del Odio. [10] hace una comparación con tres clasificadores (*BLR classifier*, *RF Decision Tree* y *SVM*) y luego establece un meta-clasificador conformado por una combinación de los tres anteriores para tomar la decisión final de clasificación. Mientras que otros autores hacen uso de algoritmos de clasificación estadísticos como *Bootstrapping* [77] y *Naïve Bayes* [62]. Por último, [48] hace uso del aprendizaje profundo con Redes Neuronales Recurrentes para la clasificación.

A la hora de extraer los conjuntos de entrenamiento y de prueba existen múltiples fuentes adoptadas por los trabajos relacionados con el enfoque de comentarios ofensivos en las RRSS. Twitter[22] [82] [83] [13] [14] [10] [68] [53], YouTube[23] [26], Yahoo[24] [57] [27] [76], Formspring[25] [26], ask.fm[26] [75], Xanga[27] [19] y Whisper[28] [68].

Prevención del discurso del odio

Además de la propia detección del discurso de odio la anticipación al mismo observando el clima social propiciado por una serie de eventos plantea un enfoque preventivo de incidentes desencadenados por el odio, como la violencia racial, ataques terroristas, asaltos por razón de pertenencia a un colectivo, etc.

En [13] estudian la posibilidad de pronosticar picos de tensión social a través de las RRSS utilizando técnicas de Minería de Textos y Análisis de Sentimientos. Y en [79] investigan el fenómeno de odio ocurrido inmediatamente después de un atentado terrorista. En esta casuística también se basa [52] para describir una taxonomía de la comunicación violenta en la red propiciada por el atentado terrorista ocurrido en París contra la revista satírica "*Charlie Hebdo*".

4.2.3 Análisis de Sentimientos

El campo de investigación conocido como Análisis de Sentimientos o Minería de Opiniones se centra principalmente en la polaridad de las expresiones plasmadas en texto escrito incluyendo aquella sin ningún sentimiento implícito. Según [43] [44] las frases que expresan sentimiento suelen ser frases subjetivas ya que entiende que los sentimientos son hechos inherentemente subjetivos.

Para entender qué es un sentimiento u opinión [43] [44] lo define como una quíntupla:

$$(e, a, s, h, t) \tag{4.1}$$

donde:

[22] https://twitter.com
[23] https://www.youtube.com
[24] https://yahoo.com
[25] https://spring.me
[26] ask.fm
[27] Xanga
[28] http://whisper.sh

- e es la entidad objetivo.
- a es el elemento objetivo de la entidad e sobre la cual se ha emitido el juicio.
- s es el sentimiento de la opinión sobre el elemento a de la entidad e. s puede ser positivo, negativo o neutro, o incluso una clasificación más granular.
- h es el emisor de la opinión.
- t el momento en el que se establece la opinión.

Por ejemplo: "Antonio a 06/01/2019: La pantalla de tu nuevo ordenador se ve realmente bien", la quíntupla sería tal que así: *ordenador*, *pantalla*, *sentimiento positivo*, *Antonio*, 06/01/2019.

Dentro del campo del Análisis de Sentimientos existen múltiples técnicas de aplicación para la clasificación de sentimientos, en [47] presenta una categorización refinada de dichas técnicas (Ver Figura 4.2)

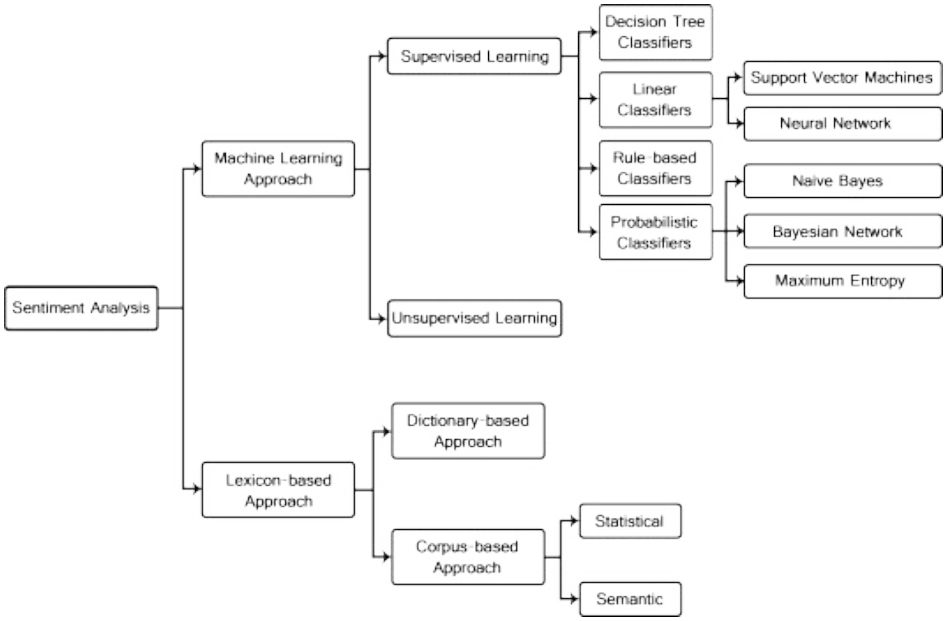


Figura 4.2: Técnicas de clasificación para el Análisis de Sentimientos [47].

Análisis de Sentimientos para la detección del discurso del odio

Es lógico asumir que la polaridad de los mensajes está estrechamente relacionada con la detección del discurso del odio, ya que se puede inferir que de un mensaje de odio se desprende un sentimiento negativo.

No es de extrañar por tanto, que varios autores hagan uso de esta técnica como soporte para la detección del odio. En el caso de [26] utilizan el Análisis de Sentimiento como paso previo a la clasificación del discurso estableciendo la polaridad de cada palabra, [70] sigue el mismo enfoque etiquetando su conjunto de datos como positivo o negativo con diferente grado de pertenencia y en [32] detectan la subjetividad de las oraciones y eliminan las frases objetivas con el fin de valorar la polaridad de esas oraciones restantes creando un modelo que a partir de la subjetividad y de características semánticas sean capaces de

identificar el odio en la red. Existen enfoques en los que usan el Análisis de Sentimiento como características para el proceso de aprendizaje supervisado como es el caso de [75] en el que emplea cuatro características numéricas de las que tres de ellas son el número de ocurrencias de palabras positivas, negativas y neutras que hay en el texto y el cuarto parámetro que indica la polaridad promedia de los textos.

Existen herramientas que facilitan la detección de la polaridad, por ejemplo, clasificadores que determinan la polaridad y además el grado de intensidad, como es el caso de SentiStrength[73] que es usado por [13], o recursos léxicos como SentiWordNet[29] que asigna a cada "*synset*" extraído de WordNet la polaridad y objetividad, este recurso es usado por [70] para la detección de frases objetivas para su posterior eliminación. En [36] hace una recopilación y análisis detallado de quince servicios de libre acceso para el Análisis de Sentimientos.

4.2.4 Sistemas Basados en el Conocimiento

Los Sistemas de ayuda a la toma de decisiones o Sistemas Expertos son programas que tratan de emular el conocimiento en una disciplina específica de un humano experto para la resolución de un problema, con el objetivo de disponer del proceso de razonamiento de un experto en cualquier parte. Los pasos fundamentales para el desarrollo de un sistema experto son los siguientes:

- Adquisición del conocimiento
- Representación del conocimiento
- Motor de inferencias

Existen pocos trabajos en los que hacen uso del conocimiento de expertos para identificar el odio. En el trabajo de [26] parten de que aplicar estereotipos normalmente atribuidos a las mujeres, al género masculino, se puede inferir una intención de insultar al destinatario. Otro ejemplo de trabajo basado en el conocimiento es el de [53] que a partir de la clasificación manual hecha en [52] que se basa en una serie de criterios criminológicos para identificar la comunicación violenta y de odio, se usaron como entrenamiento para el clasificador propuesto en su trabajo.

4.2.5 Lógica Borrosa

El concepto de conjunto borroso (Fuzzy Set) fue propuesto por primera vez en 1965 por Lofti A. Zadeh [84], definido como:

"dado un conjunto universal X, un conjunto borroso A de X se caracteriza por una función de pertenencia $f_A(X)$ la cual asocia a cada punto de X un valor entre $[0, 1]$."

Esto supuso el nacimiento de lo que conocemos actualmente como Lógica Borrosa.

La lógica borrosa (entre otros usos) permite establecer y modelar etiquetas lingüísticas a características específicas dependiendo de la situación (universo de discurso).

[29] https://sentiwordnet.isti.cnr.it/

Los juristas no se ponen de acuerdo en cómo medir la intensidad/gravedad del discurso del odio fruto de ello es la ambigüedad interpretativa que expresa el artículo 510 del CP, la lógica borrosa puede ayudar a etiquetar cómo de grave es el discurso del odio en la red atendiendo a diferentes parámetros.

CAPÍTULO 5
METODOLOGÍA

En este capítulo se expone la metodología de trabajo empleada para el desarrollo del proyecto, donde se justificará la elección de la metodología escogida y su aplicación en el marco del proyecto.

5.1 METODOLOGÍA DE DESARROLLO

Dada la naturaleza del proyecto al que podríamos calificar como un sistema de ayuda a la toma de decisiones sin entrar en la definición canónica de este tipo de sistemas, se parte de un problema mal estructurado dada la matriz terminológica del discurso del odio y su evolución interpretativa. La identificación del mismo no sólo se puede basar en la incitación directa al odio, existen factores subjetivos que hay que tener en cuenta y presentan un reto en el modelado de los mismos, por tanto requiere una revisión gradual en el proceso de desarrollo para establecer una taxonomía del odio sólida teniendo en cuenta el contexto en el que se produce el discurso.

Se torna imprescindible hacer uso de una metodología que proporcione una serie de criterios para identificar, ordenar y evaluar el proceso de desarrollo. Dado que el objetivo del proyecto no es únicamente el desarrollo de un software, requiere inicialmente un proceso de investigación y definición de una taxonomía. Es difícil establecer una metodología de desarrollo software tradicional, ya que imponen una disciplina de trabajo que requiere una firme especificación de requisitos donde los objetivos del proyecto deben estar claramente definidos.

Por el contrario una metodología ágil permite el desarrollo de un software mediante un ciclo iterativo e incremental con una planificación adaptativa según las necesidades del proyecto. En el manifiesto ágil[1] se puede observar una comparativa sobre las ventajas que tiene frente a una metodología de desarrollo tradicional:

- «**Individuos e interacciones** sobre procesos y herramientas.»
- «**Software funcionando** sobre documentación extensiva.»
- «**Colaboración con el cliente** sobre negociación contractual.»
- «**Respuesta ante el cambio** sobre seguir un plan.»

[1] http://agilemanifesto.org/iso/es/manifesto.html

Dentro de las metodologías ágiles existen diferentes "*frameworks*" de trabajo para llevar a cabo esta filosofía de desarrollo tales como **Scrum** cuyo objetivo es la obtención de resultados rápidos lo que permite una mayor adaptabilidad a las necesidades o cambios propuestos por el cliente o **Kanban** cuyo objetivo es permitir visualizar el trabajo, determinar el límite del trabajo en curso y medir el tiempo que se tarda en realizar una tarea.

Por tanto, para el desarrollo del proyecto se ha optado por un desarrollo iterativo e incremental dividiendo el proyecto en iteraciones, lo que dará como resultado final un prototipo funcional que formará el corazón de un nuevo sistema al que en un trabajo futuro se le podrán añadir mejoras y requisitos adicionales. Dado que el proyecto tiene una componente investigadora las primeras iteraciones no darán como resultado un software funcional si no el conocimiento extraído y los algoritmos diseñados.

Antes de comenzar con el desarrollo del proyecto es necesario que todos los miembros involucrados en el mismo estén en sintonía, para ello se hará uso de "***Agile Inception Deck***"[61] compuesto de una serie de dinámicas que permiten ayudar a conducir el proyecto en la dirección adecuada. Lo que se conoce por "*Inception*[2]" está formado por diez preguntas para conseguir el objetivo antes mencionado. Las preguntas a responder son las siguientes:

1. ***¿Por qué estamos aquí?***

   Esta cuestión permite contextualizar el proyecto respondiendo el por qué del mismo.

2. ***El "Elevator pitch"***

   Conocido como el discurso del ascensor, trata de responder a las preguntas; qué, por qué y para qué; utilizando el símil del tiempo que dura un viaje en el ascensor.

3. ***Diseñar una caja para el producto***

   Reflexión del producto desde las perspectiva del cliente.

4. ***Lista de los NO***

   Qué no es el producto y qué no cubre.

5. ***Conoce a tus "vecinos"***

   Conocer las personas involucradas en el proyecto.

6. ***Muestra la solución***

   Dar a conocer la metodología y la arquitectura a usar en la construcción del proyecto.

7. ***¿Qué nos quita el sueño por las noches?***

   Identificar los posibles imprevistos que puedan surgir durante el desarrollo del proyecto.

8. ***Tamaño del proyecto***

   Planificación a alto nivel de la duración del proyecto.

9. ***Muestra con claridad lo que se va a dar***

   Identificar los temas más importantes en el desarrollo del proyecto.

[2] https://agilewarrior.wordpress.com/2010/11/06/the-agile-inception-deck/

10. ***Muestra lo que va a conllevar***
    Análisis de costes del proyecto.

CAPÍTULO 6

DESARROLLO

En el presente capítulo se ilustrará todo el proceso de desarrollo del prototipo para la detección del Discurso del Odio, siguiendo la dinámica que proporciona *"Agile Inception"* para el inicio del proyecto y continuando con el desarrollo iterativo e incremental para la primera versión del sistema computacional.

Para la resolución del proyecto han sido necesarias seis iteraciones para cubrir los objetivos propuestos en el Capítulo §2 (Ver figura 6.1).

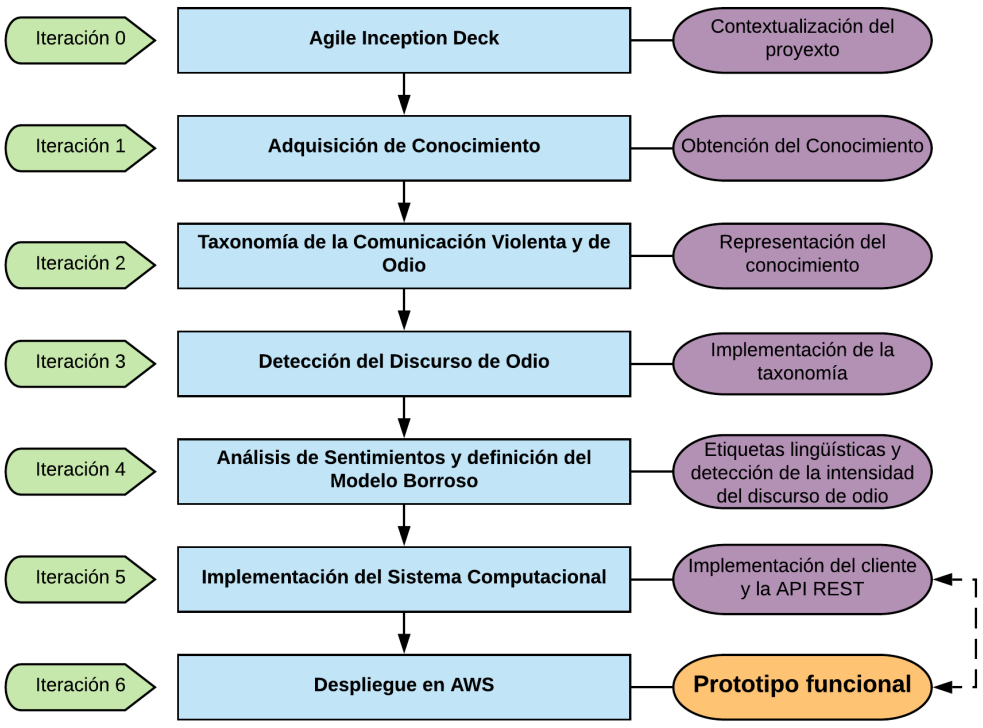


Figura 6.1: Iteraciones necesarias para llevar a cabo el prototipo para la detección del Discurso de Odio.

A lo largo del capítulo se mostrarán ejemplos explícitos de Discurso de Odio para ilustrar el desarrollo. La finalidad del proyecto es académica e investigadora en ningún momento se pretende ofender al lector. Para paliar la lacra del contenido de odio en la red es necesario conocerlo.

6.1 AGILE INCEPTION DECK

Dado que el proyecto aúna dos disciplinas dispares (Informática y Derecho), es necesario que todas las personas involucradas en el proyecto se enfoquen hacia un mismo objetivo. Para este fin, se hace uso de la metodología "*Agile Inception Deck*"[61].

1. ***¿Por qué estamos aquí?***

   La necesidad de detección del Discurso del Odio es indiscutible, como se vio en la Sección §1.1 el auge de las Redes Sociales supuso un nuevo universo para la difusión del odio en la red.

   Los distintos países están empezando a aplicar medidas contra este tipo de comunicación y se está empezando a legislar sobre la responsabilidad que tiene los proveedores de servicios como medio de difusión masiva de mensajes.

   Por último, y no menos importante se hacen necesarios mecanismos para la detección del discurso de odio en diferentes idiomas, ya que como se ha visto en el Capítulo §4, la mayoría de los estudios que hacen uso de técnicas de IA y PLN están en inglés.

2. ***El "Elevator Pitch"***

   En esta fase de la dinámica de "*Agile Inception*" consiste en imaginar que estás en el ascensor con un posible cliente y en treinta segundos hay que explicar la finalidad del proyecto, para ello se definió el prototipo como:

   Sistema computacional de Análisis de Sentimientos para la detección del Discurso del Odio haciendo uso de una taxonomía. Para llegar a dicho objetivo se emplearán técnicas de Procesamiento de Lenguaje Natural, Análisis de Sentimientos y Lógica Borrosa con el fin de, además de detectar este tipo de comportamiento delictivo identificar la intensidad del mismo (Ver Figura 6.2).

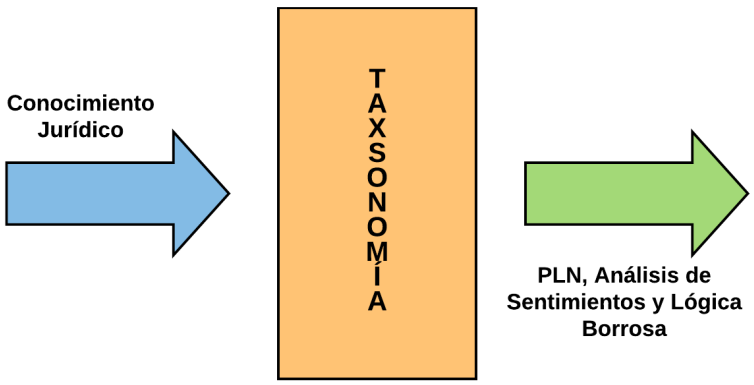


Figura 6.2: Diagrama resumen del proyecto.

3. ***Diseñar una caja para el producto***

   El prototipo parte de la necesidad de identificar el Discurso de Odio contra un colectivo diana determinado. Si suponemos un producto definitivo en el que el bien jurídico protegido sea todo colectivo social amparado por el artículo 510 del CP, la imagen ideal de la finalidad del producto sería:

- Ayudar a los proveedores de servicios de RRSS (Twitter, Facebook, etc) a la identificación de Comentarios Violentos y de Odio que se publiquen en dichas plataformas, para que apliquen sus programas de responsabilidad social corporativa y decidan qué hacer con ese tipo de mensajes. Y en el caso de condena su retirada inmediata.
- Ayudar a moderadores de foros y sección de comentarios en periódicos a identificar este tipo de mensajes.
- Ayudar a juristas y fuerzas y cuerpos de seguridad del estado a conocer la intensidad del Discurso de Odio y a jueces y fiscales a la toma de decisiones.

El prototipo limita la imagen ideal a un grupo social amparado por el art. 510 del CP, pero la finalidad es la misma.

4. ***Lista de los NO***

   Dado que es un prototipo se definieron ciertas limitaciones para acotar el proyecto:

   - Alcance del prototipo:
     - El prototipo será diseñado para un solo colectivo diana amparado por el art. 510 del CP.
     - El idioma establecido para el análisis de potenciales Comentarios de Comunicación Violenta y de Odio es el castellano.
     - Detección de incitación e injurias contra el colectivo diana y comportamientos agravantes sobre este tipo de conductas.
     - Establecimiento de etiquetas lingüísticas y mecanismo para calcular la intensidad del mensaje.
     - Implementación del prototipo funcional teniendo en cuenta los puntos antes mencionados.
   - Fuera del alcance:
     - Cubrir los colectivos que protege el art. 510 del CP en su totalidad, ya que para el prototipo se elegirá uno concreto.
     - Extracción automática de mensajes de las RRSS haciendo uso de sus APIs para ser analizados.

5. ***Conoce a tus "vecinos"***

   Este apartado trata de establecer los actores implicados en el proyecto. Los componentes del proyecto son los siguientes:

   - Autor del proyecto:
     - Andrés Montoro Montarroso.
   - Directores del Proyecto:
     - Dr. José Ángel Olivas Varela.
     - Dr. Adán Nieto Martín (experto).
   - Organizaciones/grupos implicados:

- Grupo de investigación SMILe[1]
- Instituto de Derecho Penal Europeo e Internacional[2]

6. ***Muestra la solución***

   En este apartado se muestra un resumen de la arquitectura del proyecto a alto nivel (Ver Figura 6.3) y las herramientas empleadas para el desarrollo del mismo.

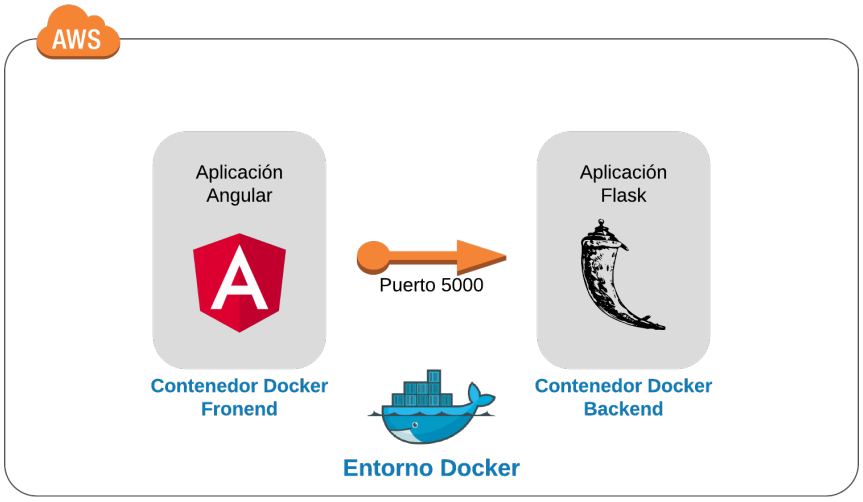


Figura 6.3: Arquitectura del proyecto.

Para llevar a cabo el proyecto se han empleado las siguientes herramientas:

- **Python 3**[3] como lenguaje de programación para la implementación de los algoritmos.
- **Freeling 4.1**[4] como librería para le Procesamiento de Lenguaje Natural.
- **SciKit-Fuzzy**[5] para la implementación del modelo borroso.
- **Docker**[6] para el despliegue de la aplicación.
- **AWS** para el despliegue de un entorno de producción (el servidor proporcionado por AWS se encuentra alojado en Irlanda).
- **Angular 7**[7] para el desarrollo de la parte de cliente.
- **Flask**[8] como framework de Python 3 para el desarrollo de la API REST.
- **Latex** para la documentación del proyecto con **Overleaf**[9] como editor online del documento (la plantilla de Latex empleada para el desarrollo del proyecto ha sido creada por el profesor Jesús Salido Tercero[10]).

[1] http://webpub.esi.uclm.es/investigacion/grupos/smile
[2] http://institutoderechopenal.uclm.es
[3] https://www.python.org/download/releases/3.0/
[4] https://github.com/TALP-UPC/FreeLing/tree/v4.1
[5] https://pythonhosted.org/scikit-fuzzy/overview.html
[6] https://www.docker.com
[7] https://angular.io
[8] https://flask-restful.readthedocs.io/en/latest/
[9] https://www.overleaf.com/
[10] http://www.uclm.es/profesorado/jsalido

- **proto.io**[11] para el diseño de la interfaz de cliente.
- **Lucidchart**[12] para el desarrollo de diagramas explicativos dentro de la documentación del proyecto.
- Repositorio git privado proporcionado por el servicio **CodeCommit**[13] de AWS.
- **Redmine**[14] para la gestión de tareas internas.

7. ***¿Qué nos quita el sueño por las noches?***

   Existen dos grandes problemáticas a la hora de llevar a cabo el proyecto:

   - la primera es definir una taxonomía para la interpretación del Delito de Odio. Dado que la matriz terminológica del Discurso del Odio presenta grandes dificultades en su interpretación, se hace imprescindible tener en cuenta los conocimientos del experto y estudios monográficos sobre el Discurso del Odio para establecer un corolario interpretativo del discurso;
   - la segunda y no menos importante es medir la intensidad del discurso del odio ya que el art. 510 del CP expresa una ambigüedad interpretativa. Para ello se hará uso del conocimiento del experto y de la Lógica Borrosa para establecer una aproximación lo más precisa posible a la intensidad del Discurso del Odio.

8. ***Tamaño del proyecto***

   A priori no se conoce la duración exacta del proyecto ya que la ambigüedad y dificultad que se plantea en el apartado anterior impide una aproximación precisa a la finalización del proyecto. Conocido esto se lanza una propuesta a alto nivel de la duración del proyecto dividido en iteraciones como se expone en el Capítulo §5 (ver Figura 6.4).

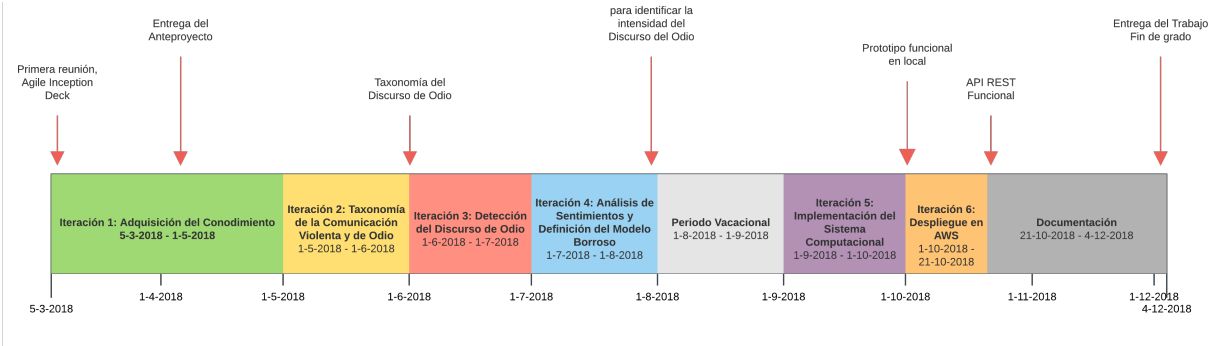


Figura 6.4: Linea temporal del Trabajo Fin De Grado.

Dado que el proyecto está en el marco del Trabajo de Fin de Grado de la Universidad de Castilla-La Mancha, está supeditado a las fechas de entrega establecidas por la Escuela Superior de Informática de Ciudad Real.

9. ***Muestra con claridad lo que se va a dar***

   Los objetivos a cumplir son los siguientes:

[11] https://proto.io
[12] https://www.lucidchart.com
[13] https://aws.amazon.com/es/codecommit/
[14] https://www.redmine.org

- Establecimiento del dominio del proyecto.
- Adquisición del conocimiento.
- Definición de la taxonomía del Discurso del Odio.
- Definición de patrones para la detección de CVydO.
- Establecimiento de etiquetas lingüísticas mediante Análisis de Sentimientos.
- Diseño del modelo Borroso para conformar la intensidad del Discurso de Odio.
- Representación funcional del sistema computacional con cliente Angular para una cómoda interacción con el prototipo.
- Despliegue de la infraestructura del servicio en Amazon Web Service[15]

10. ***Muestra lo que va a conllevar***

    El análisis de costes del proyecto se desarrolla en base a la planificación mostrada en el punto 8 (ver Figura 6.4).

    Los principales costes del proyecto se pueden dividir en dos categorías: coste de personal y coste de infraestructura.

    Los costes de personal se basan en el rol que se toma como desarrollador del proyecto que comprenderá labores de ingeniería del conocimiento, analista y programador al que se ha estipulado un coste por hora aproximado de 20€/h, y la labor del experto como especialista genuino en derecho penal al que se ha atribuido un coste por hora de 40€/h. La distribución de horas aproximadas asociadas a cada una de las iteraciones del proyecto se pueden ver en la Tabla 6.1.

    Por último, el coste de la infraestructura está comprendido por los servicios necesarios a contratar en AWS, a saber: AWS Fargate para Amazon ECS[16] con 1 vCPU a 0,04048$/h y 2 GB de memoria a 0,004445$/h por GB en la región de Irlanda. Cabe destacar que el proceso de selección de la instancia ECS de AWS no ha supuesto un proceso exhaustivo de investigación ya que para el prototipo se buscaba una máquina que permitiese un uso eficiente para una cómoda interacción con el sistema.

    En la Tabla 6.2 se muestra un desglose de los costes del proyecto.

6.2 ITERACIÓN 1: ADQUISICIÓN DEL CONOCIMIENTO

Una vez establecido el tema a tratar en el proyecto y el alcance del mismo, se procede a desarrollar cada una de la iteraciones que lo componen.

En la primera iteración se ilustra el proceso de adquisición del conocimiento lo que dará una visión más detallada del proyecto y establecerá las bases para el desarrollo de la taxonomía para la identificación de los Comentarios Violentos y de Odio.

[15] https://aws.amazon.com
[16] https://aws.amazon.com/es/fargate/pricing/

Etapas	Horas dedicadas por el Autor	Horas dedicadas por el Experto
Iteración 0: Agile Inception Deck	8h	8h
Iteración 1: Adquisición del conocimiento	156h	12h
Iteración 2: Taxonomía de la Comunicación Violenta y de Odio	92h	8h
Iteración 3: Detección del Discurso de Odio	84h	2h
Iteración 4: Análisis de Sentimientos y Definición del Modelo Borroso	88h	8h
Iteración 5: Docker, Flask y Angular	80h	1h
Iteración 6: Despliegue en AWS	60h	2h
Documentación	128h	0h
Total	**696h**	**41h**

Tabla 6.1: Horas del proyecto.

Concepto	Desglose	Horas	Coste
Personal	Autor del proyecto	696h	13.920€
	Experto	41h	1.640€
Infraestructura	AWS Fargate para Amazon ECS	456h	19,78€ aprox
Total	**15.579,78€**		

Tabla 6.2: Costes del proyecto.

6.2.1 Discurso del Odio: definición y alcance del tipo penal

Existen múltiples preceptos en el Código Penal español cuyo bien jurídico protegido es el ciudadano objetivo de conductas de odio. Los que protegen los Delitos de Odio en sentido estricto son los artículos 510 (ver Anexo A), 510 bis, 22.4º (ver Anexo A) y el art. 514.4º que versa sobre asociaciones ilícitas con motivo de odio. Luego existen múltiples artículos con una visión más amplia y omnicomprensiva como son el art. 511 del CP donde se castiga a funcionarios públicos por denegación de servicios en base a motivos de odio, el art. 512 CP cuyo bien jurídico protegido es el mismo que en el precepto anterior donde la discriminación parte del ejercicio profesional o empresarial, el art. 170.1º CP de amenazas por razón de pertenencia a una minoría social, el art. 197.5º CP por delitos de descubrimiento y revelación de secretos que afecten a datos de carácter personal en el que revelen, religión, creencias, salud, origen racial, vida sexual, menor de edad o discapacitada, etc.

Con la ayuda del experto se estableció que el precepto principal cuya propuesta interpre-

tativa está relacionada directamente con el Discurso de Odio es el art. 510 del CP (ver Anexo A) en concreto los preceptos complementarios 1 y 2, y como agravante del tema a tratar el 3. Ver Tabla 6.3.

Tabla 6.3: Artículo 510.1.a, 510.2.b y 510.3 del CP.

1. Serán castigados con una pena de prisión de uno a cuatro años y multa de seis a doce meses:
a) **Quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, hostilidad, discriminación o violencia contra un grupo, una parte del mismo o contra una persona determinada** por razón de su pertenencia a aquél, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad.
2. Serán castigados con la pena de prisión de seis meses a dos años y multa de seis a doce meses:
a) **Quienes lesionen la dignidad de las personas mediante acciones que entrañen humillación, menosprecio o descrédito de alguno de los grupos a que se refiere el apartado anterior, o de una parte de los mismos, o de cualquier persona determinada por razón de su pertenencia a ellos** por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad, o produzcan, elaboren, posean con la finalidad de distribuir, faciliten a terceras personas el acceso, distribuyan, difundan o vendan escritos o cualquier otra clase de material o soportes que por su contenido sean idóneos para lesionar la dignidad de las personas por representar una grave humillación, menosprecio o descrédito de alguno de los grupos mencionados, de una parte de ellos, o de cualquier persona determinada por razón de su pertenencia a los mismos.
3. Las penas previstas en los apartados anteriores se impondrán en su mitad superior **cuando los hechos se hubieran llevado a cabo a través de un medio de comunicación social, por medio de internet o mediante el uso de tecnologías de la información, de modo que, aquel se hiciera accesible a un elevado número de personas.**

Una vez identificado el concepto de Discurso de Odio dentro del marco jurídico penal español, se hace necesario contextualizar el precepto, para ello se propone el Plan de Acción de Rabat (ver Anexo F), propuesta que organiza mediante un test las conductas de incitación al odio para poder establecer un criterio de intensidad frente al delito. Este test ha sido sintetizado por la Recomendación de Política General número 15 (ECRI) [28]. En la tabla 6.4 se ilustra dicho test.

Cabe destacar que aparte de la ayuda indispensable del experto, la recopilación y el estudio del estado del arte (ver Capítulo §4) han sido imprescindibles para la comprensión

Tabla 6.4: Plan de Acción del Rabat (Anexo F)

(a) "el contexto en el que se utiliza el discurso de odio en cuestión (especialmente si ya existen tensiones graves relacionadas con este discurso en la sociedad)";
(b) "la capacidad que tiene la persona que emplea el discurso de odio para ejercer influencia sobre los demás (con motivo de ser por ejemplo un líder político, religioso o de una comunidad)";
(c) "la naturaleza y contundencia del lenguaje empleado (si es provocativo y directo, si utiliza información engañosa, difusión de estereotipos negativos y estigmatización, o si es capaz por otros medios de incitar a la comisión de actos de violencia, intimidación, hostilidad o discriminación)";
(d) "el contexto de los comentarios específicos (si son un hecho aislado o reiterado, o si se puede considerar que se equilibra con otras expresiones pronunciadas por la misma persona o por otras, especialmente durante el debate)";
(e) "el medio utilizado (si puede o no provocar una respuesta inmediata de la audiencia como en un acto público en directo)";
(f) "la naturaleza de la audiencia (si tiene o no los medios para o si es propensa o susceptible de mezclarse en actos de violencia, intimidación, hostilidad o discriminación)".

del Discurso de Odio y así poder proporcionar una propuesta interpretativa fundamentada en relación con el art. 510 del CP.

6.2.2 Colectivo Diana

Dado que la matriz interpretativa del precepto es muy amplia y el bien jurídico protegido hace referencia a múltiples colectivos, esto ha llevado a tomar la decisión de establecer un único **Colectivo Diana** (grupo social sobre el que se cometen los actos de odio).

De entre los diversos colectivos o grupos sociales que pueden resultar amparados por el art. 510 del CP, para el desarrollo del prototipo se identificarán mensajes de odio dirigidos contra el colectivo árabe y/o musulmán[17]. Existen dos razones por las cuales actualmente es generado un gran peligro en el que dicho colectivo sea objeto de delitos de odio. En primer lugar, la inmigración ilegal, una parte significativa de los ciudadanos extranjeros que se encuentran en nuestro país, ya sea con residencia legal o no legal, pertenecen a este colectivo. El rechazo que en determinados sectores de la población genera la inmigración se centra en mensajes contra este colectivo. En segundo lugar, y obviamente por razones de terrorismo. No es infrecuente que en el discurso de odio anti-islam se mezclen ambos argumentos. En la figura 6.5 se observa cómo en el año 2016 la mayoría de Delitos de Odio registrados son de racismo y/o xenofobia.

[17]Se habla de árabes, musulmanes e islamistas de manera indistinta, no se pretende aunar en el mismo grupo a las personas que se sientan identificadas con uno o más de estos calificativos.

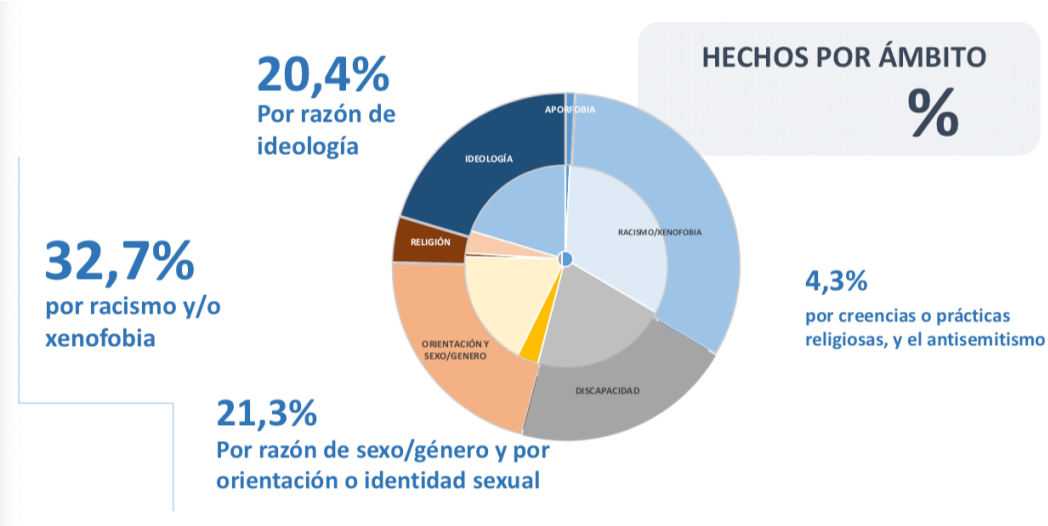


Figura 6.5: Motivos de comisión de Delitos de Odio en 2016 [66].

6.2.3 Experimento

Como última fase de adquisición de conocimiento se diseñó un experimento con el objetivo de obtener mensajes de odio en los que se recoja una amplia variedad de estructuras lingüísticas y expresiones propias del dominio.

Para ello se propuso un experimento a los alumnos de la Facultad de Derecho y Ciencias Sociales de la Universidad de Castilla-La Mancha en el campus de Ciudad Real, en el que se especifican una serie de escenarios/situaciones con el objeto de obtener un conjunto heterogéneo de mensajes, pero dentro de un dominio delimitado. Los supuestos son los siguientes:

- Escribe un comentario en el que animes a un número indeterminado de personas a ejercer actos de violencia contra el colectivo musulmán en general.
- Escribe un comentario en el que animes a un grupo concreto de personas a que realicen actos de violencia contra un grupo de musulmanes muy determinados (por ejemplo, una familia de musulmanes que viven en tú calle).
- Escribe un comentario en el que animes a un número indeterminado de personas a que discriminen al colectivo musulmán en un determinado espacio (por ejemplo, la sanidad, la educación, el empleo…).
- Escribe un comentario en el que sin mencionar palabras relacionadas con la violencia o la discriminación, menosprecies en general a las personas pertenecientes al colectivo musulmán o islamista, tachándolos a todos por ejemplo de terroristas o de personas que reciben excesivas ayudas por parte del Estado.
- Igual que en el apartado anterior pero en relación a un grupo determinado de musulmanes (por ejemplo, una familia de tu barrio …).
- Escribe una expresión que consideres que constituye un acto de menosprecio humillación o descrédito contra los musulmanes o islamistas en general, contra un grupo

concreto o una persona perteneciente a ese grupo por razón de su pertenencia al mismo.

El resultado del experimento se encuentra en el Anexo C. La redacción completa del experimento se puede ver en el Anexo G.

6.3 ITERACIÓN 2: TAXONOMÍA DE LA COMUNICACIÓN VIOLENTA Y DE ODIO

Una vez adquirido el conocimiento propio del dominio, el siguiente paso es la conceptualización de ese conocimiento adquirido en forma de taxonomía.

Para ello se ha de descomponer el Delito de Odio e identificar qué hace que un mensaje sea de odio.

6.3.1 Incitación e Injurias

Como se ha visto en la sección anterior el Delito de Odio está compuesto por varios preceptos, pero el núcleo de la Comunicación Violenta y de Odio se basa en dos sub-tipos:

a) ***"«Quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, hostilidad, discriminación o violencia contra un grupo, una parte del mismo o contra una persona determinada (...)»".***

b) ***"«Quienes lesionen la dignidad de las personas mediante acciones que entrañen humillación, menosprecio o descrédito de alguno de los grupos a que se refiere el apartado anterior, o de una parte de los mismos, o de cualquier persona determinada por razón de su pertenencia a ellos (...)»".***

Por tanto la identificación de un mensaje de odio pasa por detectar la intención de incitar a la violencia y/o injuriar contra el colectivo diana. En una aproximación formal inicial se puede definir el discurso de odio en estas tres sentencias:

Sea:

- D el colectivo diana.
- O el mensaje de Odio.
- I injuria.
- V incitación a la violencia.

Entonces:

$$\begin{aligned} \text{Si} \quad & I \wedge D \rightarrow O \\ \text{Si} \quad & V \wedge D \rightarrow O \\ \text{Si} \quad & (I \wedge V) \wedge D \rightarrow O \end{aligned} \tag{6.1}$$

Dentro de la definición de Delito de Odio la incitación al odio es más grave que la injuria, de hecho la injuria se agrava si favorece un clima de violencia[18] ,es decir, que sea la antesala del paso al acto de violencia. Por tanto, podemos decir que un acto injurioso no implica incitación, pero un acto de incitación puede contener injurias.

Una vez establecido el tipo básico de Discurso de Odio se pasa a la definición de agravantes del mismo ya que la intensidad del Discurso del Odio varía según el contexto y el clima en el que se produce.

6.3.2 Agravantes propios del mensaje

Como unidad mínima para la identificación de un mensaje como Discurso de Odio se requiere que haya una injuria contra un colectivo diana. La incitación se considera un agravante a la hora de identificar el mensaje de odio.

La medida principal para detectar cómo de grave es el Discurso del Odio viene dado por el influjo del mensaje en la sociedad, es decir, la capacidad que tiene el mensaje de agitar a la sociedad para el "*paso al acto*" de violencia contra colectivos por el mero hecho de pertenecer a ellos. Estos agravantes parten de una **incitación directa o indirecta a la violencia**.

Los agravantes propios del mensaje establecidos son los siguientes:

- **Focalizar la incitación en el tiempo**: Cuando cometes un acto de odio y a ese acto le otorgas una dimensión temporal aumenta la posibilidad del paso al acto. Por ejemplo "*Vamos a pegar a esos moros mañana*". Como se observa en el ejemplo ya se está expresando la intención de cuándo se va a cometer el supuesto acto de violencia.
- **Focalizar la incitación en el espacio**: La indicación explícita de dónde se va a cometer el acto de odio supone un agravante porque se precisa el lugar donde es posible que se cometa un acto de violencia por motivo de odio.
- **Focalizar la incitación en un sub-grupo o grupo pequeño del colectivo diana**: Precisar a quién o quienes son los objetivos de actos de violencia, es un agravante claro del mensaje. Ya que, por ejemplo si diriges la incitación contra una familia o un vecino de un barrio o incluso la identificación directa del individuo por el simple hecho de pertenecer a ese grupo acerca la incitación al paso al acto.
- **Animar a grupos a cometer actos de violencia contra el colectivo diana**: Esto es promover un sentimiento de odio animando a una multitud para cometer actos violentos contra un colectivo vulnerable por razones de pertenencia, por ejemplo: "*Vecinos vamos a echar a patadas a esos moros de mierda de nuestro barrio*".

[18] *«(…) Los hechos serán castigados con una pena de uno a cuatro años de prisión y multa de seis a doce meses cuando de ese modo se promueva o favorezca un clima de violencia, hostilidad, odio o discriminación contra los mencionados grupos (…) »*

Por último se añade la dimensión de "ensañamiento" contra el colectivo diana en la que **la contundencia del mensaje aumenta con el empleo de calificativos injuriosos y hostilidades reiteradas**.

6.3.3 Agravantes del entorno

El universo de discurso o contexto en el que se desarrolla el prototipo es la detección de mensajes de odio en redes y medios sociales. Este contexto ya supone un agravante del Delito de Odio propiamente dicho[19].

El test de severidad del Plan de Acción del Rabat (Ver Anexo F) especifica una serie de pasos para medir la intensidad del Discurso de Odio que han resultado de gran utilidad en complementación con el experto.

"*la capacidad que tiene la persona que emplea el discurso de odio para ejercer influencia sobre los demás (con motivo de ser por ejemplo un líder político, religioso o de una comunidad)*"

De esta sentencia en el marco de las Redes Sociales se pueden extraer los siguientes agravantes:

- **Número de seguidores** en el caso de que el Discurso de Odio se cometa en una Red Social.
- **Número de "Me gusta"** del potencial mensaje de odio.
- **La influencia del emisor del mensaje**; no es lo mismo que el mensaje lo publique un cargo público o personalidad reconocida a que lo publique un ciudadano sin aparente influencia relevante.

"*el medio utilizado (si puede o no provocar una respuesta inmediata de la audiencia como en un acto público en directo)*"

Dado que el acto de odio se comete en un medio virtual esta sentencia no se puede tomar al pie de la letra, pero es interpretable. El agravante extraído de dicha interpretación es:

- **Alcance del medio en el que se publica el mensaje de Odio**: este agravante hace referencia al número de receptores potenciales del Discurso de Odio, ya que por ejemplo la red social Twitter tiene un alcance masivo frente a otros medios sociales como podría ser la sección de comentarios de un periódico.

Por último, del siguiente punto:

"*la naturaleza de la audiencia (si tiene o no los medios para o si es propensa o susceptible de mezclarse en actos de violencia, intimidación, hostilidad o discriminación)*"

[19]Las penas previstas en los apartados anteriores se impondrán en su mitad superior cuando los hechos se hubieran llevado a cabo a través de un medio de comunicación social, por medio de internet o mediante el uso de tecnologías de la información, de modo que, aquel se hiciera accesible a un elevado número de personas.

Se puede extraer un agravante que comparte enunciado con el punto del mencionado test:

- **La naturaleza de la audiencia**: cuya interpretación parte del público susceptible a leer el mensaje de odio, como por ejemplo público mayoritariamente menor de edad.

Cabe destacar otro agravante propio del universo en el que se produce, y es **la cantidad de mensajes de odio que un usuario tiene publicado en sus redes**. Cuanta más cantidad de mensajes de odio tiene un usuario publicado en su perfil más propenso a cometer actos de violencia y a incentivarlos.

6.3.4 Clima

Por último y no menos importante, es imprescindible tener en cuenta la situación político-social en el que se produce el discurso de odio. A pesar de la subjetividad que pueda reflejar, los parámetros de clima son muy importantes para medir la intensidad del discurso.

- Si se ha producido un **atentado de manera reciente**.
- Si se ha producido una **oleada de inmigrantes reciente**.
- Si la convivencia es **normalizada o de máxima tensión**. Esto puede darse por una situación social precaria en la que se culpe a la inmigración legal o ilegal. O un clima de crispación en la que se hayan sucedido múltiples agresiones de odio en un corto periodo de tiempo. Obsérvese que esta última puede ser dependiente o influida por las dos anteriores.

Como se puede observar la principal particularidad de los agravantes propios del clima es que son dependientes del bien jurídico protegido, en el caso que nos ocupa, el colectivo árabe y/o musulmán.

6.3.5 Mapa de Conocimiento

Toda la conceptualización del proceso de razonamiento y la confluencia de toda la taxonomía se pone de manifiesto en el siguiente mapa de conocimiento (ver figura 6.6).

6.4 ITERACIÓN 3: DETECCIÓN DEL DISCURSO DEL ODIO

En esta iteración se cubre todo el proceso de PLN para la detección del Discurso de Odio en mensajes de texto. Para ello se hará uso del conocimiento extraído representado en la taxonomía del odio (ver Sección 6.3) y de la ontología a mencionar.

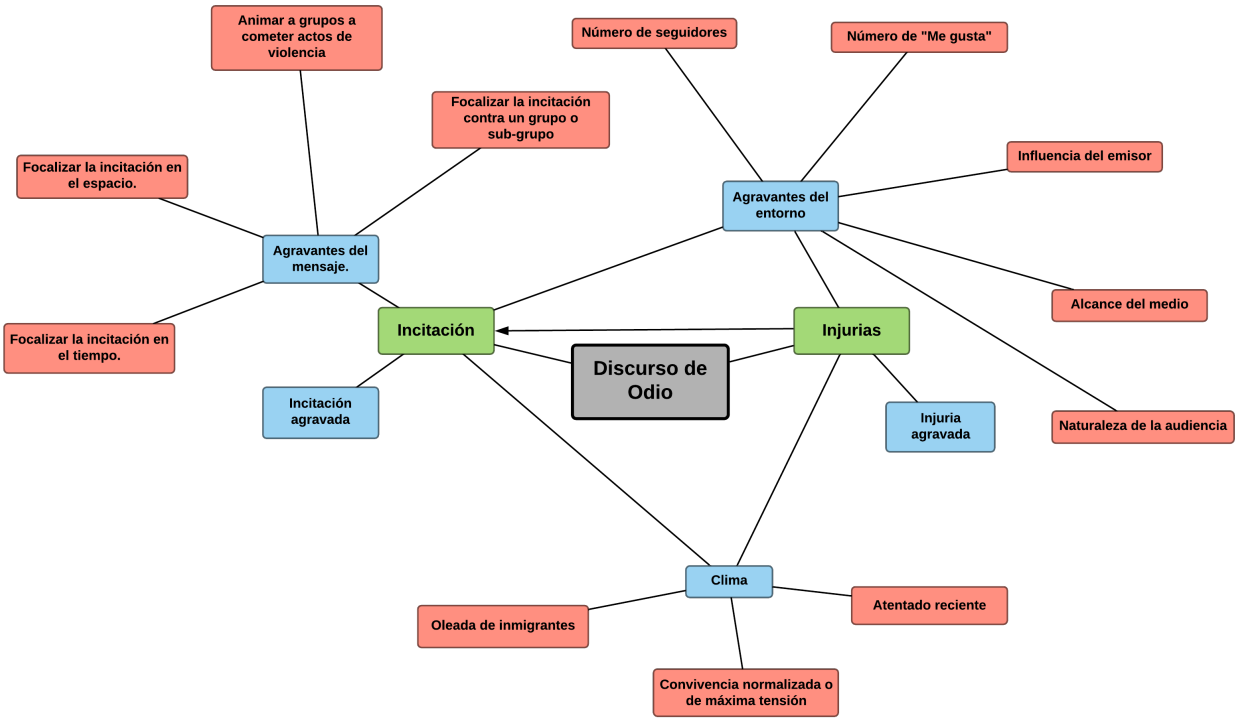


Figura 6.6: Mapa de conocimiento de la taxonomía para identificar el Discurso de Odio.

6.4.1 Ontología del dominio

Para la detección de mensajes de odio se torna imprescindible el uso de una ontología que cubra el dominio en cuestión.

Existen múltiples herramientas para el diseño de ontologías entre ellas se encuentra la herramienta protégé[20] que es una plataforma de código abierto para el diseño y construcción de ontologías haciendo uso de OWL 2[21] entre otros , que es un lenguaje ontológico para la Web Semántica. Dado que el uso de esta herramienta requería una curva de aprendizaje y las necesidades del prototipo no requieren de su potencia se optó por usar JSON (JavaScript Object Notation) que es un formato para el intercambio de datos con una sintaxis sencilla que facilita el procesamiento automático.

Para la construcción de la ontología se hizo uso del experimento lanzado a los alumnos de Derecho de la UCLM (ver Anexo C) para obtener una primera aproximación de los términos empleados en el dominio. El procedimiento para obtener los términos más relevantes de los mensajes de odio es el siguiente:

1º. Obtención de una lista con cada una de los mensajes de odio.

2º. Tokenización de la lista de mensajes en palabras.

3º. Eliminación de las "*stop words*" o palabras vacías, es decir, palabras que no aportan significado al mensaje (artículos determinados e indeterminados, adjetivos determina-

[20] https://protege.stanford.edu
[21] https://www.w3.org/TR/owl2-overview/#Syntaxes

tivos, pronombres, interjecciones, preposiciones, etc). La lista completa de palabras vacías se encuentra en el archivo de texto "spanish_stop_words_amm.txt".

4º. Creación de un espacio vectorial bidimensional en el que las columnas son la bolsa de palabras preprocesada en los pasos anteriores y las filas cada uno de los mensajes de odio (Ver listado 6.1).

Listado 6.1: Creación del espacio vectorial de los mensajes procedentes del experimento para la obtención de mensajes de odio.

```
"""
bag_of_words: bolsa de términos extraída de los mensajes
              de odio.
msg_wo_noise: listado de los mensajes de odio preprocesados
"""
def vectorizer (bag_of_words,msg_wo_noise):
    """
    Creación de una lista de listas donde la dimensión "x"
    es el número de palabras en la bolsa de términos y la
    dimensión "y" los mensajes extraídos del experimento.
    """
    vectorizer_matrix=[]
    for i in range(len(msg_wo_noise)):
        aux_list = [i*0 for i in range(len(bag_of_words))]
        vectorizer_matrix.append(aux_list)
    vectorizer_matrix = np.array(vectorizer_matrix)

    #Vectorización
    for i in range(len(msg_wo_noise)):
        for j in range(len(msg_wo_noise[i])):
            if (msg_wo_noise[i][j] in bag_of_words):
                aux_key=bag_of_words[msg_wo_noise[i][j]]
                aux=vectorizer_matrix[i][aux_key]
                aux = aux + 1
                vectorizer_matrix[i][aux_key] = aux
            else:
                vectorizer_matrix[i][j] = 0
    return vectorizer_matrix
```

5º. Obtención de la frecuencia de términos a partir del espacio vectorial.

El resultado es una lista con palabras y su frecuencia de aparición. En la tabla 6.5 se puede observar una muestra del resultado (para ver el resultado completo ir al archivo "freq_words.csv").

Del resultado del preproceso y del conocimiento plasmado en la taxonomía se ha construido una ontología con dos clases bien diferenciadas. La perteneciente al Discurso de Odio propiamente dicho y al de los agravantes del mismo, a saber:

- **Clase perteneciente al Discurso de Odio** (*hate-speech-ontology* en "ontology.json"):
  - **Colectivo diana** (*colectivo-diana* en "ontology.json"): en esta sub-clase se encuentran todos los términos relativos al colectivo diana en cuestión (árabe y/o

Tabla 6.5: Muestra de la frecuencia de términos extraída del experimento para la obtención de mensajes de odio.

Término	Frecuencia
moros	99
musulmanes	53
terroristas	31
barrio	26
mierda	25
familia	16
putos	15
islam	9
vecinos	9
…	…

musulmán), haciendo uso de la jerga discriminatoria con la que se pretende injuriar a este grupo social (moro, morito, Mustafá, etc).

  - **Injurias** (*injurias* en "ontology.json"): esta sub-clase abarca todo lo que se refiere a insultos y/o palabras malsonantes propiamente dichas.
  - **Incitación** (*incitacion* en "ontology.json"): sub-clase mayoritariamente formada por verbos que incitan a la violencia (matar, pegar, agredir, etc) y sustantivos cuyo significado transmite hostilidad (golpe, puñalada, etc).

- **Clase perteneciente a los agravantes del Discurso de Odio** (*hate-speech-aggravating-ontology* en "ontology.json"):

  - **Agravante de tiempo** (*agravante-temporal* en "ontology.json"): compuesta en su mayoría por adverbios de tiempo supliendo las carencias de la herramienta seleccionada para el Procesamiento de Lenguaje Natural (ver sub-Sección §6.4.2).
  - **Agravantes de lugar** (*agravante-lugar* en "ontology.json"): sub-clase compuesta por adverbios de tiempo (para paliar las carencias antes referenciadas) y sustantivos que indican lugar (calle, parque, vecindario, etc).
  - **Clase de grupos y subgrupos** (*agravante-grupos* en "ontology.json") : con esta subclase se pretende cubrir los agravantes de incitación contra grupo y sub-grupo por razón de pertenencia y el agravante de animar a grupos a cometer actos de violencia ya que ambas incluyen términos que pueden implicar a más de una persona o hacer referencia implícita a la misma (familia, policía, ejército, etc).

La ontología completa se encuentra en el archivo "ontology.json".

6.4.2 Freeling como herramienta para el Procesamiento de Lenguaje Natural

Freeling[22] es una librería escrita en C++ que proporciona una serie de herramientas para el Procesamiento de Lenguaje Natural en distintos idiomas (inglés, español, portugués, italiano, etc).

[22] http://nlp.lsi.upc.edu/freeling/node/1

Para la realización de este proyecto ha sido imprescindible el uso de esta librería. De entre todas las funcionalidades que proporciona se han empleado los siguientes módulos de procesamiento del lenguaje:

- **Tokenizador** ("*Tokenizer*"): para crear un vector de objetos de palabras.
- **Divisor de oraciones** ("*Sentence Splitter*"): utiliza como entrada el resultado del tokenizador hasta que recibe un identificador de final de enunciado (el punto). Después devuelve una lista de oraciones (una limitación del sistema a desarrollar es que solo analiza un enunciado en cada ejecución).
- **Analizador morfológico** ("*Morphological Analyzer*"): está compuesto por una serie de sub-módulos que permiten detectar los símbolos de puntuación, números, fechas, reconocimiento de entidades (personas, organizaciones, lugares), entre otros.
- **Etiquetado gramatical** ("*Part-of-Speech Tagger*"): asigna etiquetas gramaticales a cada uno de los términos a analizar haciendo uso del tri-grama de etiquetado de Markovian [6].
- **Parseador de dependencias basado en reglas** ("*Rule-based Dependency Parse*"): También llamado analizador de dependencias Txala [2] que crea un árbol de dependencias etiquetado con el análisis sintáctico del texto de entrada.
- **Parseador de dependencias estadístico y etiquetado de roles semánticos**("*Statistical Dependency Parser and SLR*"): una alternativa al parseador anterior el cual añade etiquetas semánticas a los argumentos de los predicados. Este analizador de dependencias se basa en [17] y el módulo de etiquetado de roles semánticos en [45].

La información detallada de la funcionalidad de estos módulos se encuentra en la documentación de la versión 4.1 de Freeling[23] que ha sido la empleada para la realización del actual prototipo para la detección del Discurso de Odio. La carga de los módulos anteriormente mencionados se hace mediante archivos de configuración. El código implementado para la configuración de las funcionalidades de Freeling se puede ver en el Anexo F y en el archivo *python* "freelingConfService.py".

Para facilitar el procesamiento del resultado del analizador de Freeling se ha seleccionado la salida basada en formato CoNLL en el que cada fila representa una palabra y cada columna el resultado de los diferentes módulos de procesamiento de lenguaje natural, para posteriormente transformarla en un *DataFrame* haciendo uso de la librería pandas[24]. Las columnas empleadas para el análisis en el proyecto actual son:

- Identificador del token (ID).
- La propia palabra (FORM).
- El lema de la palabra (LEMMA).
- El etiquetado gramatical (TAG).

[23] https://talp-upc.gitbook.io/freeling-4-1-user-manual/processing-classes/splitter
[24] https://pandas.pydata.org

- El clasificador de entidades (NEC).
- El análisis sintáctico (SYNTAX).
- El indicador de dependencias del árbol de relaciones (DEPHEAD).
- La función sintáctica (DEPREL).
- El etiquetado de Roles Semánticos (SRL).

6.4.3 Detección de la Comunicación Violenta y de Odio

Una vez obtenido el resultado del análisis proporcionado por Freeling se procede a la detección del Discurso de Odio definiendo diferentes patrones para las distintas casuísticas que presenta la identificación de odio en un mensaje.

Colectivo Diana

El primer paso para detectar si existe CVydO en un mensaje pasa por comprobar si explícitamente ese mensaje va dirigido contra el colectivo diana en cuestión, para ello se hace uso de la ontología del dominio, en concreto de la clase correspondiente al colectivo diana y se comprueba si existe alguna de las ocurrencias de dicha clase en el mensaje a analizar (el código correspondiente a esta funcionalidad se encuentra en "ontologyMatchService.py" en la función "targetMatch()").

Injurias

La detección de injurias en forma, es igual que en el apartado anterior, se hace uso de la clase de la ontología perteneciente a palabras injuriosas y se comprueba si existe una o más de una en el mensaje. Se asume el supuesto en el cual un mensaje que contenga términos que hagan referencia al colectivo diana e injurias dentro del mismo mensaje, este mensaje se califica como injurioso contra el colectivo diana (el código correspondiente a esta funcionalidad se encuentra en "ontologyMatchService.py" en la función "insultDetection()").

Incitación

En el supuesto de incitación al odio existen dos tipos, incitación directa al odio e incitación indirecta. Mientras que en el primer tipo de incitación se hace uso de palabras que explícitamente inciten a la violencia, por ejemplo: "Vamos a pegar a esos moros."; la incitación indirecta omite este tipo te términos incitatorios pero la intención sigue siendo promover la violencia, por ejemplo: "A por los moros".

Para detectar la incitación directa se hace uso de la ontología con los términos pertenecientes a la sub-clase "*incitacion*". Mientras que para la incitación indirecta se pueden dar dos casuísticas distintas:

- En el que la incitación indirecta sea sobre el colectivo diana.

$$a + por + colDiana$$

  Por ejemplo: “A por los moros”.

- En el que la incitación indirecta sea sobre un pronombre personal en tercera persona (tanto en singular como en plural) que haga referencia al colectivo diana.

$$a + por + pronPersonal$$

  Por ejemplo: “Moros, a por ellos”.

Al igual que en el caso de las injurias la incitación puede tornarse agravada si existe un repetición de los mencionados patrones, es decir, si existe hostilidad reiterada contra el colectivo diana.

El código completo se encuentra en el archivo “ontoMatchService.py” en la función “incitementDetection()”.

6.4.4 Detección de los agravantes propios del mensaje

Para calcular los agravantes propios del mensaje es condición imprescindible que exista incitación contra el colectivo diana (no es excluyente que haya injurias contra dicho colectivo).

Focalizar la incitación en el tiempo

Para la identificación de la incitación temporal se basa en la detección de tres patrones.

1º. **Que exista una fecha en el mensaje a analizar.**

2º. **Sintagma adverbial con complemento circunstancial de tiempo.**

3º. **Sintagma preposicional con complemento circunstancial de tiempo.**

En el primer patrón, Freeling en el etiquetado gramatical es capaz de identificar fechas en las oraciones, por tanto, partiendo del resultado del analizador de Freeling, si existe una etiqueta de tipo fecha (W) se considera que se ha focalizado la incitación en el tiempo. Por ejemplo: “Vamos a pegar a esos moros el **día 2 de marzo**”.

El segundo patrón consiste en identificar los sintagmas adverbiales en el mensaje a analizar y que estos tengan un complemento circunstancial de tiempo, es decir, un adverbio que indique temporalidad. Para ello nos fijaremos en la columna del análisis sintáctico para identificar los sintagmas adverbiales y el etiquetado de roles semánticos para identificar el sentido temporal del sintagma. En el momento que se cumplan estas dos condiciones se concluirá que se ha detectado el agravante temporal. Por ejemplo: “**Mañana** vamos a ir a pegar a esos moros de mierda”.

Por último el tercer patrón trata de identificar los sintagmas preposicionales cuyo complemento circustancial sea de lugar. Para ello se hará uso de la clase de la ontología de agravantes propios del mensajes, en concreto de la sub-clase "*agravante-temporal*" para que una vez identificados los sintagmas preposicionales del resultado obtenido del análisis con Freeling, se buscarán términos dentro del sintagma que hagan referencia a una localización en el tiempo. En este caso para detectar el complemento circunstancial se hace uso de la ontología porque en ciertas ocasiones Freeling no es capaz de asignarle una etiqueta de rol semántico temporal a los sintagmas preposicionales. Por ejemplo: "Vamos a pegar a esos morunos de mierda **en una semana**"

La implementación de la detección del agravante temporal se encuentra en el archivo python "aggravatingService.py" en la función "processTimeIncitement()" en la que se especifica la definición de los patrones anteriormente mencionados haciendo uso de expresiones regulares para detectarlos dentro del *DataFrame* resultado del análisis de Freeling.

Focalizar la incitación en el espacio

Como en el apartado anterior la incitación con agravante de lugar se basa en la detección de tres patrones principales:

1º. **Detección de una entidad tipo localización**, por ejemplo Madrid, Barcelona, etc.

2º. **Identificar adverbios de lugar.**

3º. **Sintagma preposicional con complemento circunstancial de lugar.**

En el primer patrón se busca que la función de clasificación de entidades (NEC) de Freeling, de como resultado una etiqueta correspondiente a una localización (B-LOC). Si se encuentra esta etiqueta en el mensaje de incitación a la violencia,se considera que focaliza dicha incitación en un lugar concreto. Por ejemplo: "Vamos a pegar a esos moros de mierda en **Madrid**".

El segundo patrón trata de identificar adverbios de lugar, obteniéndose del etiquetado gramatical (RG) con etiqueta de rol semántico de lugar (AM-LOC). Por ejemplo: "Vamos a pegar a esos moros **aquí**".

Por último, el tercer patrón consiste en identificar los sintagmas preposicionales cuyo complemento circunstancial sea de lugar. Para ello pueden darse dos casuísticas, que una vez identificados los sintagmas preposicionales el atributo de etiquetado de rol semántico identifique este complemento circunstancial (AM-LOC) o se tenga que hacer uso de la ontología de la sub-clase "agravante-lugar" porque los sustantivos comunes que expresan lugar, Freeling no es capaz de etiquetarlos. Por ejemplo: "Vamos a pegar a esos moros **en la calle**".

La implementación de la detección del agravante de lugar al igual que el agravante temporal se encuentra en el archivo python "aggravatingService.py". Dentro de la mencionada clase, la función que implementa el agravante es "processLocIncitement()" en la que se especifica la definición de los patrones anteriormente mencionados haciendo uso de expresiones regulares para detectarlos dentro del *DataFrame* resultado del análisis de Freeling.

Focalizar la incitación en un grupo o sub-grupo del colectivo diana

La incitación contra un grupo o sub-grupo (haciendo referencia también a un individuo concreto) se implementa haciendo uso de tres patrones:

1º. **Haciendo referencia a un nombre propio del colectivo diana por simple hecho de pertenecer a dicho colectivo.**

2º. **Hacer referencia a un sub-grupo perteneciente al colectivo diana.**

3º. **Si el colectivo diana está en plural la incitación esta dirigida contra un grupo del colectivo diana.**

En el primer patrón si se identifica una palabra de la ontología que hace referencia al colectivo diana como persona (B-PER)m gracias al identificador de entidades de Freeling, se convierte en incitación contra sub-grupo. Por ejemplo: "Vamos a ir a atacar a **Mustafá**".

Para la detección del segundo patrón se hace uso de la sub-clase de la ontología "agravante-grupos" y del propio colectivo diana. Si se identifica un término perteneciente a la ontología de sub-grupos y va seguido de un adjetivo calificativo (detectado en el etiquetado gramatical de Freeling) perteneciente al colectivo diana, es decir, que el propio colectivo diana actúe como adjetivo calificativo, es incitación contra sub-grupos, por ejemplo: "Vamos a agredir a esa **familia mora**". Otra casuística posible es que la palabra de la ontología que hace referencia a grupos vaya seguida de un sintagma preposicional en el que el núcleo de dicho sintagma sea un término que hace referencia al colectivo diana, por ejemplo: "Vamos a atacar a esa **familia por mora**".

Por último, si el colectivo diana a quien va dirigida la incitación a la violencia se encuentra en plural se considera incitación contra un grupo por el simple hecho de pertenecer a ese grupo social. Por ejemplo: "Vamos a pegar a esos **moros**".

La implementación de la detección del agravante de incitación dirigida contra sub-grupos o grupos se encuentra en la función "processSubgroupIncitement()" del archivo *python* "aggravatingService.py".

Animar a grupos a cometer actos de violencia contra el colectivo diana

El último agravante propio del mensaje consiste en animar a grupos a cometer actos de violencia contra el colectivo diana. Para ello los patrones que identifican este agravante son los siguientes:

1º. **Verbo que incita a la violencia en plural.**

2º. **Verbo que incita a la violencia en infinitivo precedido de la conjugación en plural del verbo "ir", "tener" o "deber".**

3º. **Sustantivos que que transmiten hostilidad precedidos del verbo "dar" más las conjugaciones del verbo "ir", "tener" o "deber".**

4º. **Verbo que incita a la violencia en infinitivo precedido de un término de la sub-clase de la ontología que representa grupos.**

5º. **Verbo que incita a la violencia en infinitivo precedido de un pronombre indefinido en plural.**

6º. **Verbo que incita a la violencia en infinitivo precedido de un pronombre personal de primera o segunda persona del plural.**

7º. **Verbo que incita a la violencia en infinitivo precedido de una entidad organización.**

8º. **Si el verbo que incita a la violencia está precedido por una sucesión de personas.**

El primer patrón se resuelve detectando el verbo incita a la violencia en plural con ayuda del etiquetado gramatical de Freeling. Por ejemplo: "**Peguen** a esos moros de manera indiscriminada".

El segundo patrón se identifica detectando el verbo ir, deber o tener (en plural) cuando hacen referencia al verbo principal (gracias al indicador de dependencias de Freeling) y este verbo principal pertenece a un verbo que incita a la violencia de la ontología. Por ejemplo: "**Vamos a pegar** a esos moros de mierda".

El tercer patrón tiene como peculiaridad el hecho de que la acción violenta se lleva a cabo por un sustantivo que indica hostilidad, la forma de detectarlo es haciendo uso de la ontología en concreto de la sub-clase "*incitacion*". Cuando la incitación sea un nombre común que apunta al verbo dar en infinitivo y este a su vez a las conjugaciones en plural de los verbos ir, tener o deber se considera animar a grupos a cometer actos de violencia contra el colectivo diana. Por ejemplo: "**Deberíamos dar una paliza** a esos moros de mierda".

En los siguientes patrones se emplea el infinitivo del que incita a la violencia con valor de imperativo por tanto debe aparecer siempre precedido de la preposición "a". Dadas estas condiciones se alienta a grupos a cometer actos de violencia cuando:

- Cuando haciendo uso de la ontología de grupos, el verbo que incita a la violencia hace referencia a dichos términos. Por ejemplo: "**Amigos a pegar** a esos moros".
- Cuando un pronombre indefinido en plural (excepto ningunos y ningunas) hace referencia al verbo que incita a la violencia. Por ejemplo: "Vamos **todas a pegar** a esos moros de mierda".
- Cuando el núcleo del sintagma nominal es un pronombre personal en primera o segunda persona del plural y hace referencia a un verbo que incita a la violencia. Por ejemplo: "**Vosotros a pegar** a esos moros."
- Cuando se detecta una sucesión de personas (identificadas por el reconocedor de entidades de Freeling) y son precedidas por el verbo de incitación a la violencia. Por ejemplo: "**Antonio, Pablo y Javier a pegar** a esos moros."
- Por último cuando el núcleo del sintagma nominal es una organización (identificada con el reconocedor de entidades) y el verbo que incita a la violencia tiene un indicador de dependencia que hace referencia a dicha organización. Por ejemplo: "**Policías a pegar** a esos moros"

La implementación de los mencionados patrones se encuentra con el resto de agravantes en el archivo "aggravatingService.py" en la función "processGroupIncitement()". Remarcar que también está contemplada la incitación indirecta en la aplicación de estos patrones.

6.5 ITERACIÓN 4: ANÁLISIS DE SENTIMIENTOS Y DEFINICIÓN DEL MODELO BORROSO

En esta iteración, a partir del conocimiento recopilado y la detección del Discurso de Odio y sus agravantes, se van a definir una serie de etiquetas lingüísticas basadas en el resultado del análisis del mensaje y ulteriormente se desglosará y ponderará la taxonomía con ayuda del experto para establecer un modelo borroso que permita conocer la intensidad del comentario de odio.

Dicho modelo borroso parte de la idea llevada a cabo en [55], en el que a partir de un estudio psicológico llamado *Affective Norms for English Words* (ANEW) [5] que proporciona una serie de valoraciones emocionales de un conjunto de sustantivos en inglés, se utiliza como base para describir un modelo borroso con cinco categorías (muy positivo, positivo, neutro, negativo y muy negativo) para establecer el sentimiento de mensajes extraídos de las Redes Sociales. Dicho artículo ha sido publicado en el congreso internacional de lógica borrosa FUZZ-IEEE 2018 y se ha implementado en la herramienta Prometeus IDS[25].

6.5.1 Análisis de Sentimientos. Establecimiento de las etiquetas lingüísticas

Para el establecimiento de etiquetas lingüísticas se van a definir una serie de reglas o condiciones que tiene que tener el mensaje para asignarle dichas etiquetas.

- Si en el mensaje solo existe mención al colectivo diana la etiqueta lingüística será: "**Posible uso peyorativo del Colectivo Diana**". Para el resto de reglas es imprescindible que exista un nombrado explícito del colectivo diana si no se detecta colectivo diana el mensaje se dejará de analizar.
- En referencia a la injuria:
  - Si existe una injuria entonces "**Comentario injurioso contra el colectivo diana**".
  - Si existe más de una injuria dentro del mensaje "**Comentario injurioso agravado contra el colectivo diana**".
  - Si no existe injuria "**No hay injurias contra el colectivo diana**".
- En referencia a la incitación:
  - Si existe un término que incita a la violencia entonces "**Comentario que incita a la violencia contra el colectivo diana**".

[25] https://prometeusgs.com/inicio/prometeus-intelligent-data-suite/

  - Si existe más de un término que incita a la violencia entonces "**Comentario agravado que incita a la violencia contra el colectivo diana**".
  - Si no existe incitación "**No existe incitación a la violencia contra el colectivo diana**".

- En referencia a los agravantes propios del mensaje. La condición de establecimiento de estas etiquetas es que exista incitación única o agravada (pudiendo existir injurias contra el colectivo diana).

  - Focalizar la incitación en el tiempo
    * Si existe incitación en el tiempo entonces: "**Focalización de la incitación en el tiempo**"
    * Si no existe incitación en el tiempo entonces: "**La incitación a la violencia no está focalizada en el tiempo**".
  - Focalizar la incitación en el espacio
    * Si existe incitación en el espacio entonces: "**Focalización de la incitación en el espacio**"
    * Si no existe incitación en el espacio entonces: "**La incitación a la violencia no está focalizada en el espacio**".
  - Incitación dirigida contra grupos o sub-grupos por el simple hecho de pertenecer al mismo.
    * Si existe incitación dirigida contra grupos o sub-grupos: "**Incitación dirigida contra grupos o sub-grupos por el simple hecho de pertenecer al mismo**".
    * Si no existe incitación en el espacio entonces: "**La incitación a la violencia no está focalizada contra grupos o sub-grupos pertenecientes al colectivo diana**".
  - Animar a grupos a cometer actos de violencia.
    * Si se incita a grupos a cometer actos de violencia: "**Incitación de grupos a cometer actos de violencia contra el colectivo diana**".
    * Si no existe incitación a grupos a cometer actos de violencia: "**No existe incitación que anime a grupos a cometer actos de violencia**".

Una vez definidas las reglas, como se ha mencionado el la definición de la taxonomía (ver Sección §6.3) existe una clara diferencia entre incitación e injurias, ya que la incitación se podría considerar un tipo agravado de la injuria. Por tanto, para la definición del modelo borroso se han definido dos universos de discurso el perteneciente a la injuria que se ha denominado Comunicación de Odio (CdO) y el perteneciente a la incitación en el cual se seguirá adoptando el término propuesto por Miró en [52], Comunicación Violenta y de Odio (CVydO). Sin olvidar que en un comentario considerado de incitación a la violencia puede existir injuria.

6.5.2 Asignación de pesos a la taxonomía del odio

En esta sección se establecen los pesos según el nivel de intensidad que pueda tener el discurso de odio y se desgranan las opciones de los agravantes del entorno y del clima:

- La Tabla 6.6 hace referencia a los pesos de la incitación y la injuria propiamente dichos.
- En la Tabla 6.7 se establecen pesos a los agravantes propios del mensaje.
- En la Tabla 6.8 se especifican y desglosan los agravantes propios del entorno y se asignan pesos según su intensidad.
- Y en la Tabla 6.9 se pondera el clima en distintos intervalos dándole valores específicos a etiquetas intermedias para después facilitar la construcción de las reglas para establecer el modelo borroso.

Variables del Discurso de Odio		
Clase	Especificación	Peso
Injurias (I)	Injuria leve (I_l)	1
	Injuria agravada (I_a)	1,5
Incitación (V)	Incitación leve (V_l)	4
	Incitación agravada (V_a)	6

Tabla 6.6: Asignación de pesos a las variables que componen el Discurso de Odio.

Agravantes propios del mensaje		
Clase	Especificación	Peso
Agravantes propios del mensaje (M)	Focalizar la incitación en el tiempo (M_t)	5
	Focalizar la incitación en el espacio (M_e)	5
	Focalizar la incitación contra grupo o sub-grupo (M_s)	6
	Animar a grupos a cometer actos de violencia (M_g)	4

Tabla 6.7: Asignación de pesos a los agravantes propios del mensaje

6.5.3 Construcción del modelo borroso

Una vez establecidos los pesos de cada una de las variables se procede al desarrollo de funciones de pertenencia de los modelos borrosos.

Como se ha mencionado en la sub-sección §6.5.1 existen dos universos de discurso claramente diferenciados en lo que se refiere al Discurso de Odio. Por tanto, se llevará a cabo la construcción de dos modelos borrosos; uno para representar el Discurso de Odio injuriante (Comunicación de Odio) y otro para representar el Discurso de Odio que incita a la violencia (Comunicación Violenta y de Odio).

Comunicación de Odio

Las funciones de pertenencia que componen este modelo borrosos serán etiquetadas de la siguiente forma:

Agravantes propios del entorno (E)		
Clase	Especificación	Peso
Número de Seguidores (E_s)	De 0 a 100 (E_{s1})	0
	De 100 a 500 (E_{s2})	1
	De 500 a 5.000 (E_{s3})	1,5
	De 5.000 a 25.000 (E_{s4})	2
	De 25.000 a 100.000 (E_{s5})	3
	Más de 100.000 (E_{s6})	5
Número de "Me gusta/compartidos" (E_g)	De 0 a 50 (E_{g1})	0
	De 50 a 250 (E_{g2})	1
	De 250 a 2.500 (E_{g3})	1,5
	De 2.500 a 10.000 (E_{g4})	2
	De 10.000 a 25.000 (E_{g5})	3
	Más de 25.000 (E_{g6})	5
Alcance del medio en el que se difunde (E_a)	Red Social Masiva (E_{a1})	2
	Sección de comentarios de un periódico (E_{a2})	1,5
	Foro en Internet (E_{a3})	1
	Otros (E_{a4})	0,5
Naturaleza de la audiencia (E_n)	Audiencia menor de edad (E_{n1})	6
	Fuerzas y cuerpos de Seguridad del Estado (E_{n2})	5
	Audiencia general (E_{n3})	0
Influencia del emisor (E_i)	Pertenece a los poderes del estado (E_{i1})	50
	Cantante y/o famoso y/o influencer (E_{i2})	14
	Autoridad religiosa (E_{i3})	7
	No relevante (E_{i4})	0

Tabla 6.8: Asignación de pesos a los agravantes propios del entorno.

- **Comunicación de Odio Leve** (CdO_{leve})
- **Comunicación de Odio Agravado** ($CdO_{agravado}$)
- **Comunicación de Odio Severo** (CdO_{severo})
- **Comunicación de Odio Muy grave** (CdO_{mgrave})

Para la representación de la función de pertenencia se hará uso de lo que se conoce como función de pertenencia trapezoidal (*Trapezoidal-shaped membership function*).

$$f(x;a,b,c,d) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c} & c \leq x \leq d \\ 0, & d \leq x \end{cases} \tag{6.2}$$

Agravantes propios del clima (E)		
Clase	Especificación	Peso
Atentado reciente (C_a) 0 - 14	Alto (C_{a1})	14
	Medio (C_{a2})	[7 - 14)
	Bajo (C_{a3})	[0 - 7)
Oleada de inmigrantes reciente (C_o) 0 - 6	Alto (C_{o1})	6
	Medio (C_{o2})	[3 - 6)
	Bajo (C_{o3})	[0 - 3)
Convivencia normalizada o de máxima tensión (C_c) 0 - 20	Alto (C_{c1})	20
	Medio (C_{c2})	[10 - 20)
	Bajo (C_{c3})	[0 - 10)

Tabla 6.9: Asignación de pesos a los agravantes propios del clima.

Donde x es un vector y a, b, c y x son parámetros escalares.

Para la representación del modelo, x representa una lista que va desde el valor más pequeño que se puede dar en la CdO hasta el mayor, es decir, la intensidad más leve del mensaje hasta la más grave teniendo en cuenta todos los parámetros que componen la CdO.

$$\begin{aligned} V_{min}CdO =& 1,5 \\ V_{max}CdO =& 109,5 \end{aligned} \tag{6.3}$$

Para la definición de los escalares de cada una de las etiquetas se han supuesto una serie de escenarios supervisados por el experto para construir la función de pertenencia.

- Comunicación de odio Leve ($Cd0_{leve}$).

$$\begin{aligned} Cd0_{leve}a =& I_l + E_{s1} + E_{g1} + E_{a4} + E_{n3} + E_{i4} + C_{a3}(0) + C_{o3}(0) + C_{c3}(0) \\ =& V_{min}CdO = 1,5 \\ Cd0_{leve}b =& I_l + E_{s1} + E_{g1} + E_{a4} + E_{n3} + E_{i4} + C_{a3}(0) + C_{o3}(0) + C_{c3}(0) \\ =& V_{min}CdO = 1,5 \\ Cd0_{leve}c =& I_l + E_{s3} + E_{g3} + E_{a3} + E_{n3} + E_{i4} + C_{a3}(2) + C_{o3}(1) + C_{c3}(3) \\ =& 11 \\ Cd0_{leve}d =& I_l + E_{s4} + E_{g4} + E_{a1} + E_{n3} + E_{i4} + C_{a3}(4) + C_{o3}(2) + C_{c3}(4) \\ =& 17 \end{aligned} \tag{6.4}$$

- Comunicación de odio Agravado ($Cd0_{agravado}$).

$$\begin{aligned} Cd0_{agravado}a &= I_l + E_{s2} + E_{g2} + E_{a1} + E_{n3} + E_{i4} + C_{a3}(4) + C_{o3}(2) + C_{c3}(3) \\ &= 14 \\ Cd0_{agravado}b &= I_l + E_{s4} + E_{g4} + E_{a1} + E_{n3} + E_{i4} + C_{a2}(7) + C_{o2}(3) + C_{c3}(5) \\ &= 22 \\ Cd0_{agravado}c &= I_l + E_{s5} + E_{g5} + E_{a1} + E_{n3} + E_{i4} + C_{a2}(7) + C_{o1}(6) + C_{c3}(5) \\ &= 27 \\ Cd0_{agravado}d &= I_l + E_{s5} + E_{g5} + E_{a1} + E_{n3} + E_{i4} + C_{a2}(10) + C_{o1}(6) + C_{c3}(10) \\ &= 35 \end{aligned} \tag{6.5}$$

- Comunicación de odio severo ($Cd0_{severo}$).

$$\begin{aligned} Cd0_{severo}a &= I_a + E_{s1} + E_{g1} + E_{a3} + E_{n1} + E_{i3} + C_{a2}(7) + C_{o2}(3) + C_{c3}(5) \\ &= 30,5 \\ Cd0_{severo}b &= I_a + E_{s6} + E_{g5} + E_{a1} + E_{n3} + E_{i2} + C_{a2}(7) + C_{o1}(6) + C_{c3}(5) \\ &= 43,5 \\ Cd0_{severo}c &= I_a + E_{s6} + E_{g6} + E_{a1} + E_{n3} + E_{i2} + C_{a1}(14) + C_{o1}(6) + C_{c2}(10) \\ &= 57,5 \\ Cd0_{severo}d &= I_a + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i2} + C_{a1}(14) + C_{o1}(6) + C_{c2}(12) \\ &= 65,5 \end{aligned} \tag{6.6}$$

- Comunicación de odio muy grave ($Cd0_{mgrave}$).

$$\begin{aligned} Cd0_{mgrave}a &= I_a + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i2} + C_{a2}(11) + C_{o1}(6) + C_{c2}(10) \\ &= 60,5 \\ Cd0_{mgrave}b &= I_a + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i2} + C_{a1}(14) + C_{o1}(6) + C_{c2}(18) \\ &= 71,5 \\ Cd0_{mgrave}c &= I_a + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i1} + C_{a1}(14) + C_{o1}(6) + C_{c1}(20) \\ &= V_{max}CdO = 109,5 \\ Cd0_{mgrave}d &= I_a + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i1} + C_{a1}(14) + C_{o1}(6) + C_{c1}(20) \\ &= V_{max}CdO = 109,5 \end{aligned} \tag{6.7}$$

Como resultado de las funciones se obtiene el modelo borroso que se puede observar en la figura 6.7.

Comunicación Violenta y de Odio

Al igual que en el apartado anterior las funciones de pertenencia que componen este modelo borrosos serán etiquetadas de la siguiente forma:

- **Comunicación Violenta y de Odio Leve** ($CVydO_{leve}$)
- **Comunicación Violenta y de Odio Agravado** ($CVydO_{agravado}$)

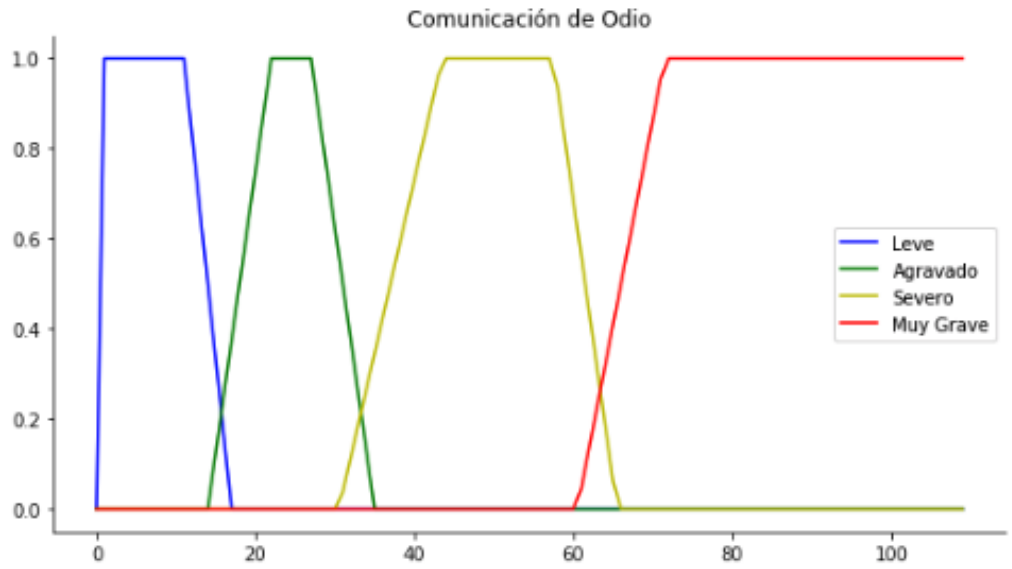


Figura 6.7: Modelo borroso de la Comunicación de Odio.

- **Comunicación Violenta y de Odio Severo** ($CVydO_{severo}$)
- **Comunicación Violenta y de Odio Muy grave** ($CVydO_{mgrave}$)

Se hará uso de la misma función trapezoidal para la representación de la función de pertenencia.

$$f(x;a,b,c,d) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c} & c \leq x \leq d \\ 0, & d \leq x \end{cases} \tag{6.8}$$

Donde x es un vector y a, b, c y x son parámetros escalares.

El vector x representa una lista que va desde el valor más pequeño que se puede dar en la CVydO hasta el máximo , es decir, la intensidad más leve del mensaje hasta la intensidad más grave al igual que en el modelado de CdO.

$$\begin{aligned} V_{min}CVydO &= 4,5 \\ V_{max}CVydO &= 135,5 \end{aligned} \tag{6.9}$$

Para la definición de los escalares de cada una de las etiquetas se han supuesto una serie de escenarios supervisados por el experto para construir la función de pertenencia.

- Comunicación de odio Leve ($CVydO_{leve}$).

$$\begin{aligned}
CVydO_{leve}a =& V_l + E_{s1} + E_{g1} + E_{a4} + E_{n3} + E_{i4} + C_{a3}(0) + C_{o3}(0) + C_{c3}(0) \\
=& V_{min}CVydO = 4,5 \\
CVydO_{leve}b =& V_l + E_{s1} + E_{g1} + E_{a4} + E_{n3} + E_{i4} + C_{a3}(0) + C_{o3}(0) + C_{c3}(0) \\
=& V_{min}CVydO = 4,5 \\
CVydO_{leve}c =& V_l + E_{s4} + E_{g4} + E_{a1} + E_{n3} + E_{i4} + C_{a3}(2) + C_{o3}(1) + C_{c3}(3) \\
=& 16 \\
CVydO_{leve}d =& V_l + E_{s4} + E_{g4} + E_{a1} + E_{n3} + E_{i4} + C_{a2}(7) + C_{o2}(3) + C_{c3}(3) \\
=& 23
\end{aligned} \tag{6.10}$$

- Comunicación de odio Agravado ($CVydO_{agravado}$).

$$\begin{aligned}
CVydO_{agravado}a =& V_l + M_g + E_{s1} + E_{g1} + E_{a4} + E_{n3} + E_{i4} + C_{a3}(0) + C_{o3}(0) + \\
& C_{c3}(0) = 8,5 \\
CVydO_{agravado}b =& V_l + M_t + E_{s3} + E_{g3} + E_{a1} + E_{n3} + E_{i4} + C_{a3}(2) + C_{o3}(1) + \\
& C_{c3}(3) = 20 \\
CVydO_{agravado}c =& V_l + M_s + E_{s4} + E_{g5} + E_{a1} + E_{n3} + E_{i4} + C_{a2}(7) + C_{o1}(6) + \\
& C_{c3}(4) = 34 \\
CVydO_{agravado}d =& V_l + M_s + E_{s4} + E_{g5} + E_{a1} + E_{n3} + E_{i4} + C_{a1}(14) + C_{o2}(4) + \\
& C_{c2}(10) = 45
\end{aligned} \tag{6.11}$$

- Comunicación de odio severo ($CVydO_{severo}$).

$$\begin{aligned}
CVydO_{severo}a =& I_l + V_l + M_t + E_{s1} + E_{g1} + E_{a3} + E_{n1} + E_{i3} + C_{a2}(7) + C_{o2}(3) + \\
& C_{c3}(5) = 39 \\
CVydO_{severo}b =& I_l + V_l + M_s + E_{s6} + E_{g5} + E_{a1} + E_{n3} + E_{i2} + C_{a2}(7) + C_{o1}(6) + \\
& C_{c3}(5) = 53 \\
CVydO_{severo}c =& I_a + V_l + M_t + M_e + M_s + M_g + E_{s2} + E_{g5} + E_{a1} + E_{n3} + E_{i4} + \\
& C_{a1}(14) + C_{o1}(6) + C_{c2}(10) = 61,5 \\
CVydO_{severo}d =& I_a + V_a + M_t + M_e + M_s + M_g + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i2} + \\
& C_{a1}(14) + C_{o2}(4) + C_{c3}(6) = 83,5
\end{aligned} \tag{6.12}$$

- Comunicación de odio muy grave ($CVydO_{mgrave}$).

$$\begin{aligned}
CVydO_{mgrave}a =& V_l + M_t + M_e + M_s + M_g + E_{s6} + E_{g6} + E_{a1} + E_{n3} + E_{i2} + \\
& C_{a1}(14) + C_{o3}(0) + C_{c2}(10) = 74 \\
CVydO_{mgrave}b =& V_a + M_t + M_e + M_s + M_g + E_{s6} + E_{g6} + E_{a1} + E_{n3} + E_{i2} + \\
& C_{a1}(14) + C_{o1}(6) + C_{c2}(18) = 91,5 \\
CVydO_{mgrave}c =& I_a + V_a + M_t + M_e + M_s + M_g + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i1} + \\
& C_{a1}(14) + C_{o1}(6) + C_{c1}(20) = V_{max}CVydO = 135,5 \\
CVdO_{mgrave}d =& I_a + V_a + M_t + M_e + M_s + M_g + E_{s6} + E_{g6} + E_{a1} + E_{n1} + E_{i1} + \\
& C_{a1}(14) + C_{o1}(6) + C_{c1}(20) = V_{max}CVydO = 135,5
\end{aligned} \tag{6.13}$$

Como resultado de las funciones se obtiene el modelo borroso que se puede observar en la figura 6.7.

Figura 6.8: Modelo borroso de la Comunicación Violenta de Odio.

6.6 ITERACIÓN 5: IMPLEMENTACIÓN DEL SISTEMA COMPUTACIONAL

Una vez implementados todos los algoritmos que componen el sistema computacional de Análisis de Sentimientos para la detección del Discurso de Odio, se procede al desarrollo de una aplicación web que permita la interacción con el sistema de una forma sencilla y amigable para dar por finalizado el prototipo.

La anteriormente nombrada aplicación web será implementada en Angular 7 que es un *framework* de desarrollo web. La conexión del cliente y servidor se hará mediante el *micro framework* Flask haciendo uso de servicios REST. Por último, ambas partes del proyecto (cliente y servidor) estarán contenidas cada una en un contenedor Docker que permite el despliegue de la aplicación en cualquier entorno independientemente del sistema operativo.

6.6.1 Implementación del cliente Angular

Angular es un framework de desarrollo web creado por Google cuya finalidad es crear aplicaciones SPA (*Single Page Application*) es decir, aplicaciones web que se cargan en una sola página y que se van recargando dinámicamente según el usuario interactúa con la misma.

Una de las principales ventajas de Angular es el uso de componentes que permiten realizar una división lógica del código. Angular usa como lenguaje de programación **TypeScript** que es un superconjunto de JavaScript que añade tipados y la posibilidad de crear clases.

Antes de comenzar con la implementación se realizaron una serie de diseños previos para definir una interfaz (ver Figura 6.9) que permita al usuario configurar los parámetros necesarios para el análisis del discurso del odio. Los parámetros configurables son los siguientes:

- Agravantes propios del entorno (número de seguidores, número de "Me gusta", naturaleza de la audiencia, influencia del emisor y alcance del medio) con un selector para poder configurar los metadatos asociados al mensaje a analizar.
- Los agravantes propios del clima (oleada de inmigrantes, atentado reciente y estado de la convivencia) con un selector en forma de *slider* para que el analista valore la situación actual y configure adecuadamente las respuestas planteadas en la configuración del clima.

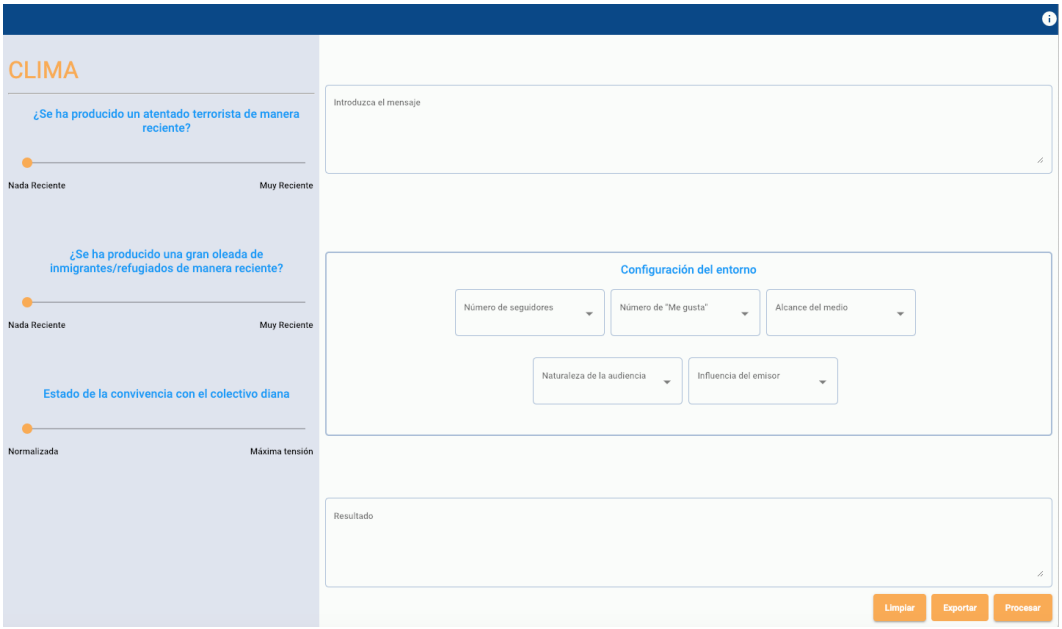


Figura 6.9: Diseño final de la interfaz de usuario.

La evolución del diseño de la interfaz se puede ver en el Anexo I.

La estructura de la aplicación parte del módulo principal llamado "*AppModule*" compuesto por varios componentes, a saber:

- ***"shade-body"***([26]): que contiene la entrada de texto donde se introducirá el mensaje potencialmente de odio a analizar y la configuración de los agravantes del entorno en forma de desplegable para seleccionar la configuración correspondiente (ver Figura 6.10).
- ***"shade-nav"***: es el componente que contiene toda la configuración con respecto a los agravantes de clima. Ya que los agravantes de clima tiene amplia granularidad en función del estado de la sociedad actual, se ha optado por una representación en forma de "*slider*" para facilitar al usuario la configuración del mismo (ver Figura 6.11).

Además a nivel lógico se incluyen los siguientes servicios y utilidades:

- ***"core"***: incluye los servicios que permiten la comunicación entre los componentes y el servidor:

[26]La palabra "shade" se encuentra en numerosas partes del código tanto en cliente como en servidor y hace referencia al nombrado del sistema cuyas siglas significan "Service of hate speech detection"

Figura 6.10: Ejemplo de selector.

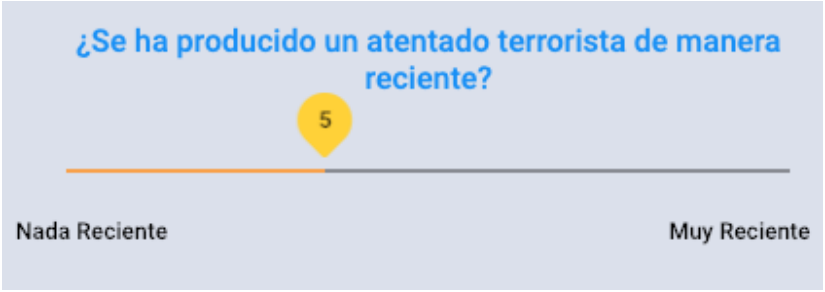


Figura 6.11: Ejemplo de "slider".

  - ***"EnvironmentAggravatingService"***: donde se realizan las peticiones http al servidor.
  - ***"ComponentCommunicationService"***: que contiene los valores seleccionados por los usuarios en los distintos formularios. Este componente es inyectado en *"shade-nav"* y en *"shade-body"* para obtener los valores antes mencionados. También es inyectado en *"EnvironmentAggravatingService"* para componer los valores y enviarlos al servidor.

- ***"shared"***: contiene las utilidades que pueden ser compartidas por los diferentes componentes. En el prototipo actual solo contiene los *"dialogs"* que muestran información al usuario (ver Figura 6.12).

Todos los componentes y servicios mencionados se encuentran alojados en la carpeta *"frontend"* del código fuente del proyecto.

Arquitectura Angular

La arquitectura de la parte de cliente esta basada en un Docker multietapa ("Docker multi-stage builds"[27]), es decir, un contenedor Docker cuya construcción se basa en dos etapas (ver Listado 6.2):

[27] https://docs.docker.com/develop/develop-images/multistage-build/

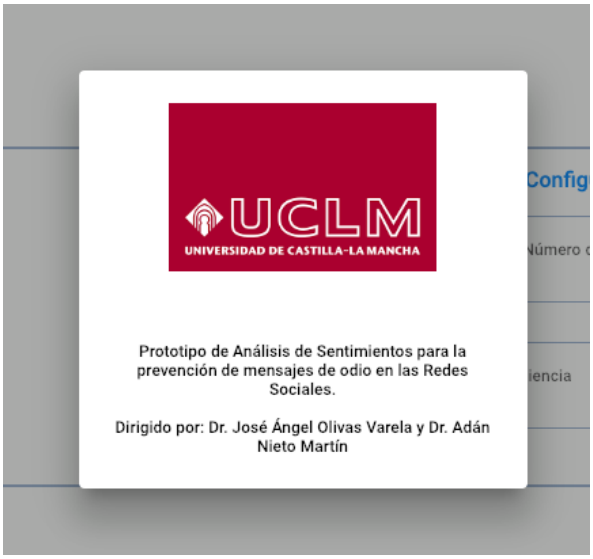


Figura 6.12: Información adicional mostrada en la interfaz.

- Una primera fase construye una aplicación Angular con una imagen base de Node.js[28] donde se construye, instala y compila todo lo necesario para la aplicación en local.
- Una segunda fase donde se crea una imagen con el servidor web NGINX[29] con el código compilado habiendo descartado utilidades específicas de Node.js para dar lugar a una aplicación eficiente lista para producción.

Listado 6.2: Dockerfile para el despligue del contenedor del cliente.

```
1
2  FROM node:10.15.0-slim as builder
3  COPY package*.json ./
4  RUN npm set progress=false && npm config set depth 0 && npm
       cache clean --force
5  RUN npm i && mkdir /ng-app && cp -R ./node_modules ./ng-app
6  WORKDIR /ng-app
7  COPY . .
8  RUN $(npm bin)/ng build --prod --build-optimizer
9
10 FROM nginx:1.14.2-alpine
11 COPY nginx/default.conf /etc/nginx/conf.d/
12 RUN rm -rf /usr/share/nginx/html/*
13 COPY --from=builder /ng-app/dist /usr/share/nginx/html
14 CMD ["nginx", "-g", "daemon off;"]
```

6.6.2 Comunicación entre cliente y servidor: Flask

Para permitir la integración en cliente del sistema computacional para la detección de delitos de odio se implementa una API REST en Flask con el objetivo de permitir peticiones sobre el servicio.

[28] https://nodejs.org/en/
[29] https://www.nginx.com

Flask es un *micro framework* de desarrollo de APIs en Python, pero no está preparado para el despliegue en producción ya que no tiene un servidor por defecto. Para paliar este problema se hará uso de NGINX y uWSGI[30].

NGINX es un servidor web y proxy inverso que usa un enfoque asíncrono basado en eventos donde cada petición se maneja en un solo hilo, esto permite que con un solo proceso principal se puedan controlar múltiples procesos de forma concurrente lo que evita bloquear otras peticiones y lo convierte en un servidor ideal para aplicaciones en producción. Las peticiones generadas por NGINX pasarán a uWSGI que es un servidor de aplicaciones capaz de comunicarse con el servidor web (NGINX) del que recibe peticiones que son enviadas a la aplicación Flask (ver Figura 6.13). El uso de Docker facilita la instalación del entorno.

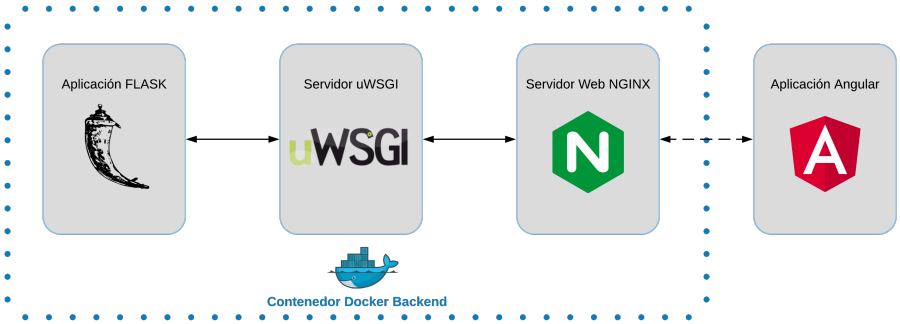


Figura 6.13: Arquitectura del servidor.

El contenedor de Docker virtualiza una maquina Ubuntu 18.04 con Python 3 y el gestor de paquetes pip3 instalado con el que se instalarán el resto de librerías para el desarrollo del proyecto[31] incluida la instalación y compilación de la librería Freeling y el anteriormente mencionado servidor de aplicaciones uWSGI, mientras que NGINX se instala mediante el gestor de paquetes "apt-get". El archivo Dokerfile donde se compone el contenedor se encuentra en la carpeta backend del proyecto.

El servicio REST implementado para el desarrollo de la aplicación recibe peticiones http del cliente para procesar la información y devolver una respuesta. La operación implementada para establecer esta comunicación es:

- Operación POST (ver Listado 6.3) para procesar el mensaje de odio junto con la configuración establecida por el usuario (agravantes del entorno y de clima).

Composición de la aplicación y despliegue en local

Para el despliegue de los contenedores en local se hace de ***Docker Compose***[32] que es una herramienta para definir y ejecutar aplicaciones multicontenedores. Con Docker Compose se emplea un archivo de configuración YALM para conformar los servicios de la aplicación (ver Listado 6.4). El archivo YALM para la configuración del despliegue de la aplicación se encuentra en el archivo "docker-compose.yml" del proyecto.

[30] https://uwsgi-docs.readthedocs.io/en/latest/
[31] Las librerías se encuentran el el archivo requeriments.txt de la carpeta backend del proyecto
[32] https://docs.docker.com/compose/

Listado 6.3: Petición POST para el procesamiento del mensaje de odio.

```
@app.route('/process',  methods=['POST'])
def processMessage():
    result = {}

    try:
        followers = int(request.args.get('followers'))
        likes = int(request.args.get('likes'))
        reach = int(request.args.get('reach'))
        public = int(request.args.get('public'))
        influence = int(request.args.get('influence'))
        terrorist_attack =
            int(request.args.get('terrorist_attack'))
        wave_immigration =
            int(request.args.get('wave_immigration'))
        living = int(request.args.get('living'))
        hate_speech = request.args.get('hate_speech')

        result = Shade().processCVydO(followers, likes,
            reach, public, influence, terrorist_attack,
            wave_immigration, living, hate_speech)
    except:
        result={"Error":"No se ha podido procesar los
            datos. Comprueba que los campos estén
            seleccionados correctamente"}
        return  jsonify(result)

    return jsonify(result)
```

Para poder recibir peticiones se expone el puerto 5000 del *container* de *backend* con el de la maquina donde se ejecuta.

Una vez configurado el archivo de Docker Compose para construir y ejecutar la aplicación en local hay que ejecutar las sentencias que se pueden encontrar en el listado 6.5.Es condición indispensable tener instalado Docker[33] y Docker Compose[34] en la máquina donde se vaya a ejecutar el prototipo.

En el archivo "Makefile" de la carpeta raíz del proyecto se encuentran todas las sentencias necesarias para el despliegue del proyecto.

6.7 ITERACIÓN 6: DESPLIEGUE EN AWS

Amazon Web Service es un servicio de computación en la nube que proporciona infraestructuras TI escalables. Para el desarrollo de este proyecto se ha hecho uso de dicha infraestructura para el despliegue del sistema computacional.

De los múltiples servicios que proporciona AWS para el desarrollo del proyecto se emplea **ECS** (Elastic Container Service[35]) que es un servicio de organización de contenedores de

[33] https://docs.docker.com/install/
[34] https://docs.docker.com/compose/install/
[35] https://aws.amazon.com/es/ecs/

Listado 6.4: Archivo de configuración YALM para el despliegue del prototipo.

```
1
2  services:
3    frontend:
4      image: shade-frontend
5      build:
6        context: ./frontend
7      container_name: "shade-frontend"
8      ports:
9        - "80:80"
10
11   backend:
12     image: shade-backend
13     build:
14       context: ./backend
15     container_name: "shade-backend"
16     ports:
17       - "5000:5000"
```

Listado 6.5: Sentencias para el despliegue del prototipo.

```
1  docker-compose -p shade up -d
```

alta escalabilidad y rendimiento compatible con Docker.

Este servicio incluye también el servicio **AWS Fargate**[36] que autogestiona las instancias que aprovisionan y ajustan los *clusters* de máquinas virtuales lo que permite enfocarse en la creación y ejecución de aplicaciones dejando a un lado la gestión de la infraestructura.

Despliegue del servicio ECS

A continuación se muestra un resumen del proceso de despliegue del servicio Amazon ECS[37]:

1. Creación del repositorio de las imágenes que componen el proyecto con ECR.
2. Subida de las imágenes de Docker (ver Listado 6.6) al repositorio creado (frontend y backend)[38].
3. Creación del *cluster* para ejecutar las imágenes. En este proceso se crea una tarea con los *containers* del proyecto (ver Figura 6.14).

Los *clusters* del servicio ECS son específicos de cada región (Irlanda en el proyecto actual). Las tarea creada utiliza el tipo de lanzamiento Fargate que como se ha mencionado

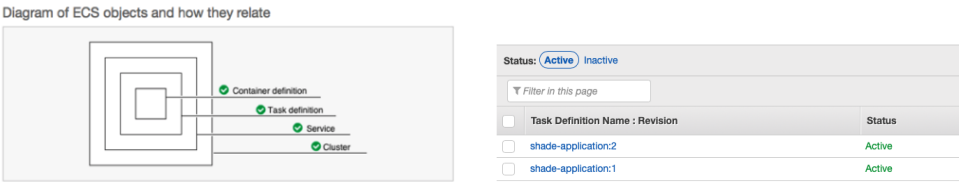
[36] https://aws.amazon.com/es/fargate/

[37] Para el proceso de despliegue en Amazon ECS se ha seguido el siguiente tutorial: https://linuxacademy.com/blog/amazon-web-services-2/deploying-a-containerized-flask-application-with-aws-ecs-and-docker/

[38] Como pre-requisito para realizar esta subida es la instalación en el equipo de AWS CLI y la configuración de las credenciales de la cuenta en AWS https://docs.aws.amazon.com/es_es/cli/latest/userguide/cli-chap-install.html

Listado 6.6: Instrucciones a ejecutar para la subida de las imágenes Docker al repositorio ECR.

```
$(aws ecr get-login --no-include-email --region eu-west-1)
docker build -t shade-backend:latest ./backend
docker build -t shade-frontend:latest ./frontend
docker tag shade-backend:latest
id-aws-user.dkr.ecr.eu-west-1.amazonaws.com
    /shade:shade-backend
docker tag shade-frontend:latest
id-aws-user.dkr.ecr.eu-west-1.amazonaws.com
    /shade:shade-frontend
docker push id-aws-user.dkr.ecr.eu-west-1.amazonaws.com
    /shade:shade-backend
docker push id-aws-user.dkr.ecr.eu-west-1.amazonaws.com
    /shade:shade-frontend
```



(a) Creación del cluster.

(b) Revisión de la definición de la tarea con los dos contenedores.

Figura 6.14: Imagen del proceso de creación del cluster para el despliegue del servicio.

permite ejecutar las aplicaciones en contenedores sin tener que aprovisionar y administrar la infraestructura del *backend*. La arquitectura general de AWS Fargate se puede ver en la figura 6.15[39].

Demo del prototipo

Para cerrar el desarrollo se muestra un ejemplo de ejecución del prototipo. Para ello se establece la siguiente configuración:

- **Mensaje de odio**: "Vamos a pegar a esos moros de mierda hoy en el parque."
- **Agravantes del Entorno**:
  - **Número de seguidores**: 5.000 - 25.000
  - **Número de "Me gusta"**: 250 - 2.500
  - **Alcance del medio**: Red Social Masiva
  - **Naturaleza de la audiencia**: Audiencia general
  - **Influencia del emisor**: No relevante
- **Agravantes de Clima**:

[39]La imagen ha sido extraída de https://docs.aws.amazon.com/es_es/AmazonECS/latest/developerguide/launch_types.html

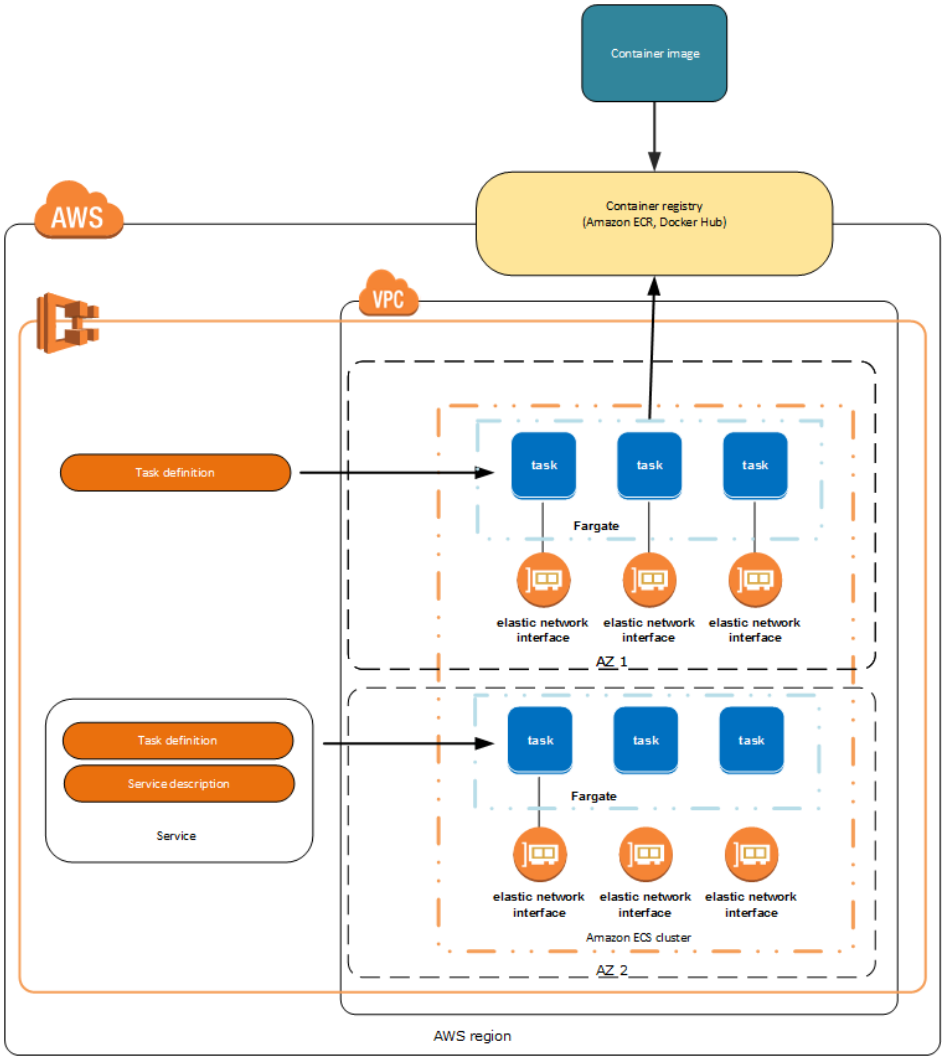


Figura 6.15: Arquitectura de AWS Fargate.

- **¿Se ha producido un atentado de manera reciente?**: 4
- **¿Se ha producido una gran oleada de inmigrantes/refugiados de manera reciente?**: 4
- **Estado de la convivencia con el colectivo diana**: 6

Ver Figura 6.16 para la mencionada configuración en la interfaz gráfica.

En la figura 6.17 se muestra el análisis del mensaje de odio expresado en el ejemplo.

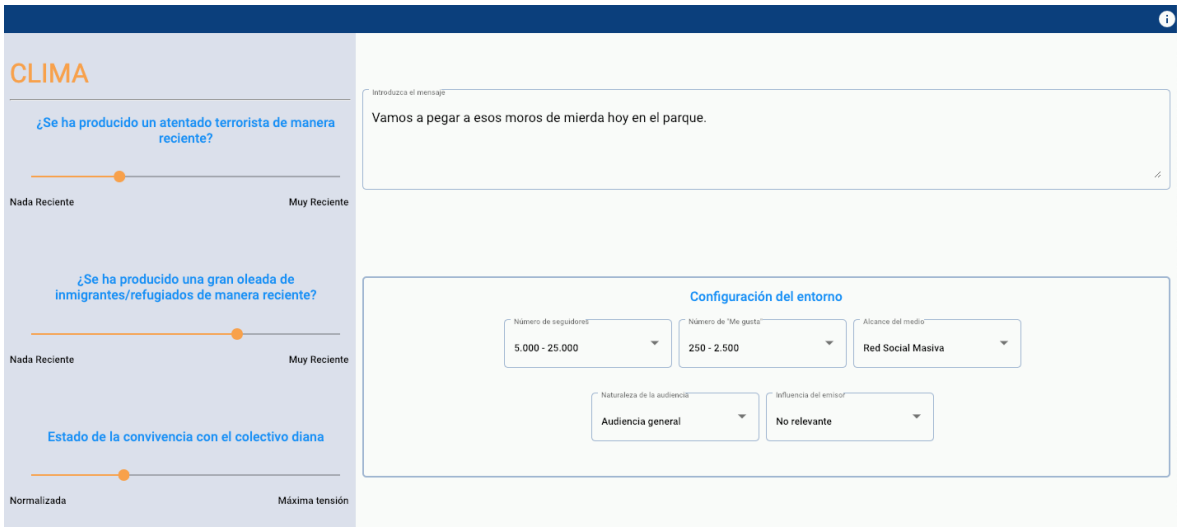


Figura 6.16: Mensaje a analizar y configuración de los agravantes.

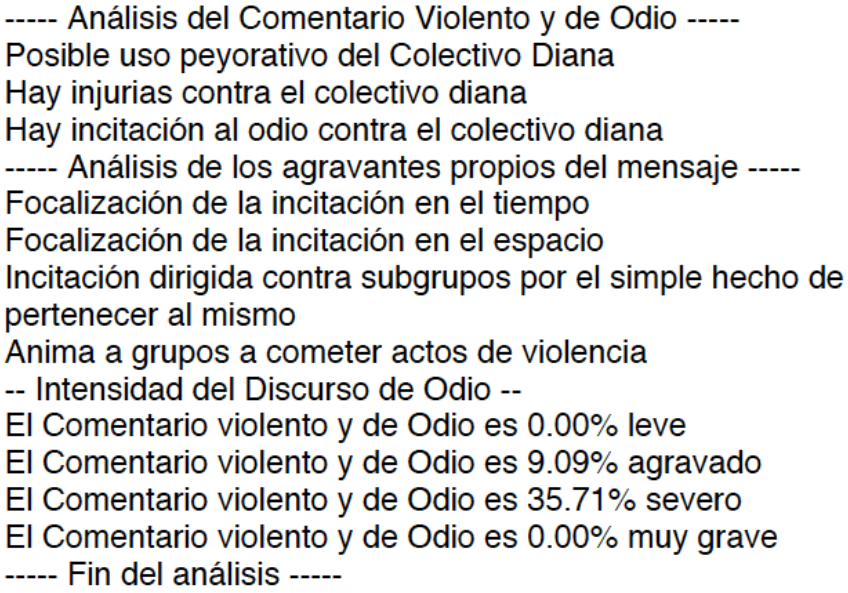


Figura 6.17: Resultado del análisis del mensaje de odio.

CAPÍTULO 7
EVALUACIÓN Y RESULTADOS

En este capítulo se muestra la evaluación del sistema a partir de los mensajes extraídos del experimento realizado en la facultad de Derecho de la UCLM en el campus de Ciudad Real. También se muestra una revisión de la duración del proyecto y por ende de los costes que conlleva el mismo.

7.1 EVALUACIÓN DEL PROTOTIPO DE ANÁLISIS DE SENTIMIENTOS PARA LA PREVENCIÓN DE MENSAJES DE ODIO EN LAS REDES SOCIALES

Partiendo de los 259 mensajes extraídos del experimento realizado con los alumnos de Derecho de la UCLM se ha lanzado una prueba sobre el prototipo donde se han analizado dichos mensajes (ver Tabla 7.1).

Evaluación		
% Aciertos	% Fallos	% Falsos Positivos respecto a Fallos
93,05 %	6,95 %	66,7 %

Tabla 7.1: Evaluación del prototipo.

Cabe destacar que los falsos positivos detectados en la Tabla 7.1 corresponden a mensajes en los que se expresa incitación a la violencia por parte del colectivo diana y no viceversa. Por ejemplo: "Los **moros** únicamente **quieren** reconquistar España a base de poner bombas y **matarnos a todos**."

7.2 ACTUALIZACIÓN DE LA PLANIFICACIÓN DEL PROYECTO

Como se vio en la sección §6.1 la aproximación a la planificación inicial establecía como fecha de finalización del proyecto el 5 de Diciembre de 2018 (coincidiendo con la convocatoria especial de finalización del la Escuela Superior de Informática de la UCLM en el campus de

Ciudad Real), dicha planificación sufre modificaciones (ver Figura 7.1) debido principalmente a la curva de aprendizaje de la librería para PLN Freeling y al establecimiento de patrones para la detección de los agravantes propios del mensaje.

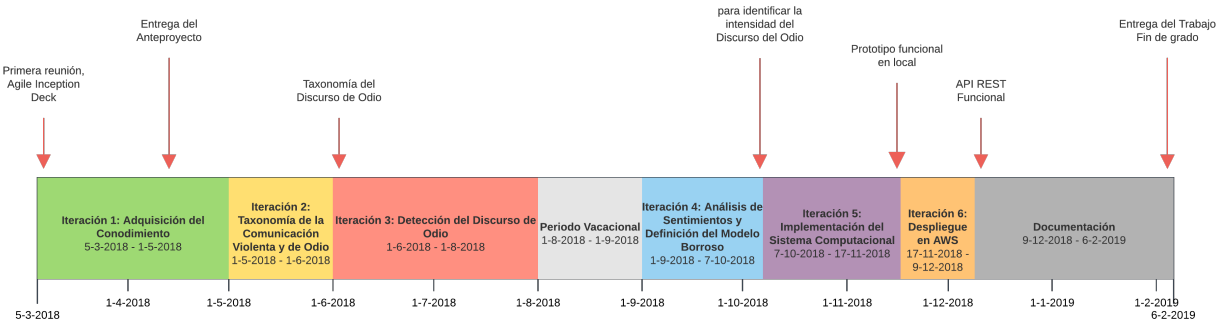


Figura 7.1: Linea temporal actualizada del Trabajo Fin De Grado.

Como consecuencia de la actualización de la planificación, la distribución de horas aproximadas asociadas a cada una de las iteraciones del proyecto y por ende los costes de desarrollo. El cómputo total de horas y el coste de desarrollo del proyecto pueden verse en la Tabla 7.2.

Concepto	**Desglose**	**Horas**	**Coste**
Personal	Autor del proyecto	876h	17.520€
	Experto	41h	1.640€
Infraestructura	AWS Fargate para Amazon ECS	456h	19,78€ aprox
Total	**19.179,78€**		

Tabla 7.2: Costes actualizados del proyecto.

CAPÍTULO 8
CONCLUSIONES

A continuación se muestra una visión global del trabajo realizado durante la elaboración del proyecto junto con los objetivos cumplidos. Por último, se hará mención de las posibles mejoras al prototipo actual y del trabajo de investigación que se desprende de la realización del proyecto.

8.1 OBJETIVOS ALCANZADOS

La finalidad del proyecto consiste en la creación e implementación de una taxonomía capaz de identificar qué comentarios eran de odio y en base a distintos parámetros identificar la intensidad del mensaje. Para acotar el universo de discurso se decidió identificar un colectivo diana en concreto (árabe y/o musulmán) objeto de odio. Como se ilustra a lo largo de la memoria, el desarrollo del proyecto ha requerido una serie de pasos para completar el prototipo de análisis de sentimientos para la detección de mensajes de odio en la red. A continuación se muestra un resumen del grado de éxito de cada uno de los objetivos descritos en el capítulo §2:

- **Adquisición del conocimiento:** Esta etapa del desarrollo del proyecto se tornó compleja, ya que el desconocimiento de una disciplina como el derecho penal, obligó a estudiar gran cantidad de bibliografía, para que con ello y con el conocimiento del experto se pudiese adquirir el conocimiento suficiente para el diseño de una taxonomía con el fin de identificar un mensaje de odio. Una vez finalizado el proceso de adquisición del conocimiento se tomó la decisión de especificar el bien jurídico protegido (el colectivo árabe y/o musulmán) con el fin de facilitar el desarrollo del prototipo y que diese una visión global del precepto a cubrir pero a su vez una reducción de la complejidad a la hora de la implementación. Para la selección del colectivo diana se llegó a un consenso con el experto argumentando la situación actual de dicho colectivo y los recientes acontecimientos en tema de inmigración ilegal y terrorismo islámico.

- **Diseño de experimento para la obtención de términos del dominio:** para conocer mejor el léxico empleado en este tipo de comunicación de odio, se diseñó un experimento con el que recoger muestras de delitos de odio. Los sujetos que llevaron a cabo el experimento fueron alumnos de derecho de la UCLM del campus de Ciudad Real. Se obtuvo una muestra de 259 potenciales mensajes de odio.

- **Creación de la ontología del dominio:** a partir de los términos extraídos del experimento se establecieron las clases de la ontología y los términos que la componen. También completada con sinónimos de los términos encontrados y palabras propias del dominio. La ontología es mejorable, ya que se encuentra sesgada en cuanto a términos del experimento se refiere.

- **Definición de la taxonomía para la identificación de mensajes de odio:** una vez extraído todo el conocimiento se modeló para diseñar una taxonomía para la identificación del discurso del odio. Dentro de la taxonomía se pueden distinguir dos clases bien diferenciadas, la identificación del Discurso de Odio propiamente dicho (con la detección de incitación a la violencia e injurias) y los agravantes, divididos a su vez, en agravantes propios del mensaje, del entorno y del clima. La taxonomía representa de forma genérica la composición del discurso del odio a excepción de los agravantes de clima que son propios del dominio. El empleo de la taxonomía permite identificar cómo de grave es el comentario de odio.

- **Definición del conjunto y etiquetas borrosas para el mecanismo de Análisis de Sentimientos:** una vez definida la taxonomía se añadió una serie de pesos a cada una de las etiquetas de la taxonomía para así poder construir un modelo borroso con funciones de pertenencia de casos tipo, supervisados por el experto. Se diferenciaron dos modelos borrosos que hacían referencia a la Comunicación de Odio (discurso injurioso con o sin agravantes del entorno y del clima) y a la Comunicación Violenta y de Odio (discurso de incitación a la violencia que puede incluir injurias y agravantes propios del mensaje junto con el resto de agravantes). Esta diferenciación de modelos ha permitido establecer una serie de etiquetas lingüísticas según el nivel de intensidad del mensaje de odio acompañados de un porcentaje de pertenencia a cada uno de los modelos borrosos según la matriz de incitación o injuriosa del mensaje. Gracias a la definición del modelo borroso la taxonomía pasa a convertirse en una métrica razonablemente precisa para medir la intensidad del discurso del odio.

- **Implementación del modelo:** para la implementación del prototipo se empleo el lenguaje de programación Python, junto con la librería Freeling, pandas y re (librería para expresiones regulares) para las tareas de PLN (detección de injurias, incitación a la violencia y los agravantes propios del mensaje) y para la implementación del modelo borroso se empleó la librería para lógica borrosa SciKit-Fuzzy.

- **Despliegue e infraestructura:** una vez implementado la parte de *backend* con todas las funcionalidades se diseñó un interfaz amigable para la configuración de los parámetros de los agravantes del entorno y del clima y poder generar un resultado en la pantalla de la aplicación web y exportar un informe pdf del análisis del potencial mensaje de odio. Para llevar a cabo la conexión entre cliente y servidor se hizo uso del *microframework* Flask elegido por su sencillez a la hora de generar servicios. Todo el despliegue de la aplicación se hizo a través de Docker y se ejecuta en local o en una maquina AWS para proporcionar un servicio accesible en remoto para la API REST.

Por tanto se puede afirmar que se han alcanzado todos los objetivos propuestos.

8.2 TRABAJO FUTURO

Dado que se trata de un prototipo de sistema se entiende que es susceptible a mejoras. En ningún caso se puede afirmar que la implementación del prototipo es completa, ya que la forma en la que nos comunicamos está en constante evolución y pueden surgir nuevos patrones de CVydO. A continuación se enumeran las carencias del prototipo actual:

1. **Falta de patrones para la detección del discurso de odio y sus agravantes propios del mensaje.** Para ello se requería un aumento del conjunto de test extraído del experimento para la detección de nuevos patrones y plantear nuevos escenarios para que los mensajes cubran un espectro más amplio de incitación al odio.
2. **Optimización de la velocidad de procesamiento del análisis del mensaje de odio, ya que Freeling tarda demasiado en procesar cada mensaje.** Las posibles soluciones pasan por estudiar más la librería para una mayor optimización del código, emplear funciones lambda proporcionadas por AWS[1] cuya ejecución en memoria incrementaría el rendimiento del análisis o incluso prescindir de esta librería e implementar los procesos de PLN que requiere la detección de patrones del actual prototipo.

Aunque el prototipo presenta algunas carencias, del desarrollo del proyecto se ha desprendido una serie de trabajos a realizar en el futuro:

- **Aumento del bien jurídico protegido expandiendo el sistema al resto de colectivos diana que protege el artículo 510 del Código Penal**. Lo que implicará un rediseño de la taxonomía añadiendo nuevos agravantes de clima propuestos en cada dominio.
- **La implementación de análisis de mensajes por usuario (dentro de una red social) y usar la variable de números de comentarios de odio como agravante propio del entorno.**
- **Conectar la API REST del prototipo con la API de cualquier red social masiva para el análisis automático de mensajes.**
- **Entrenar un modelo que sirva de filtro para identificar mensajes de odio** y únicamente se emplee la taxonomía para valorar la intensidad del mensaje de odio para optimizar el proceso de análisis.
- **Mecanismo para la configuración automática de los agravantes del entorno.**

Todos estos trabajos quedan englobados en la solicitud de fondos para un proyecto de investigación científica y transferencia tecnológica 2018 de la Junta de Comunidades de Castilla-La Mancha, en la que se formará un equipo multidisciplinar que englobará a los componentes del grupo de investigación SMiLe de la Escuela Superior de Informática de Ciudad Real y al Instituto de Derecho Penal Europeo e Internacional.

La investigación en español para la detección automática de delitos de odio y más aún clasificación por intensidad es escasa, como se ha podido ver en el estado del arte la gran

[1] https://aws.amazon.com/es/lambda/

mayoría de investigaciones del tema que atañe al proyecto actual se basa en el análisis de mensajes en inglés y prácticamente ninguna proporciona una taxonomía para la detección de mensajes de odio basándose en la legalidad vigente del llamado Delito de Odio.

Por último, se lanza una propuesta de modificación de la responsabilidad penal de los proveedores ya que en el artículo 16 de la LSSI (ver Anexo B) el proveedor deberá eliminar el contenido declarado ilícito en su servicio de manera diligente. Esta expresión denota que no es necesario que elimine el contenido inmediatamente si no que valore cuánto tiempo tiene que tardar para causar el menor perjurio posible al resto de usuarios. Esto era comprensible cuando la tecnología no era capaz de aislar un suceso y necesitaba parar todo el servicio, ahora los proveedores de servicios puede eliminar o bloquear contenido declarado ilegal inmediatamente sin causar perjurio alguno al resto de usuarios en materia de libertad de expresión.

ANEXO A

LOS DELITOS DE ODIO: ARTÍCULOS 510 Y 22.4º CP

ARTÍCULO 510

1. Serán castigados con una pena de prisión de uno a cuatro años y multa de seis a doce meses:

    a) Quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, hostilidad, discriminación o violencia contra un grupo, una parte del mismo o contra una persona determinada por razón de su pertenencia a aquél, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad.

    b) Quienes produzcan, elaboren, posean con la finalidad de distribuir, faciliten a terceras personas el acceso, distribuyan, difundan o vendan escritos o cualquier otra clase de material o soportes que por su contenido sean idóneos para fomentar, promover, o incitar directa o indirectamente al odio, hostilidad, discriminación o violencia contra un grupo, una parte del mismo, o contra una persona determinada por razón de su pertenencia a aquél, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad.

    c) Públicamente nieguen, trivialicen gravemente o enaltezcan los delitos de genocidio, de lesa humanidad o contra las personas y bienes protegidos en caso de conflicto armado, o enaltezcan a sus autores, cuando se hubieran cometido contra un grupo o una parte del mismo, o contra una persona determinada por razón de su pertenencia al mismo, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, la situación familiar o la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad, cuando de este modo se promueva o favorezca un clima de violencia, hostilidad, odio o discriminación contra los mismos.

2. Serán castigados con la pena de prisión de seis meses a dos años y multa de seis a doce meses:

a) Quienes lesionen la dignidad de las personas mediante acciones que entrañen humillación, menosprecio o descrédito de alguno de los grupos a que se refiere el apartado anterior, o de una parte de los mismos, o de cualquier persona determinada por razón de su pertenencia a ellos por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad, o produzcan, elaboren, posean con la finalidad de distribuir, faciliten a terceras personas el acceso, distribuyan, difundan o vendan escritos o cualquier otra clase de material o soportes que por su contenido sean idóneos para lesionar la dignidad de las personas por representar una grave humillación, menosprecio o descrédito de alguno de los grupos mencionados, de una parte de ellos, o de cualquier persona determinada por razón de su pertenencia a los mismos.

b) Quienes enaltezcan o justifiquen por cualquier medio de expresión pública o de difusión los delitos que hubieran sido cometidos contra un grupo, una parte del mismo, o contra una persona determinada por razón de su pertenencia a aquél por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad, o a quienes hayan participado en su ejecución.

Los hechos serán castigados con una pena de uno a cuatro años de prisión y multa de seis a doce meses cuando de ese modo se promueva o favorezca un clima de violencia, hostilidad, odio o discriminación contra los mencionados grupos.

3. Las penas previstas en los apartados anteriores se impondrán en su mitad superior cuando los hechos se hubieran llevado a cabo a través de un medio de comunicación social, por medio de internet o mediante el uso de tecnologías de la información, de modo que, aquel se hiciera accesible a un elevado número de personas.

4. Cuando los hechos, a la vista de sus circunstancias, resulten idóneos para alterar la paz pública o crear un grave sentimiento de inseguridad o temor entre los integrantes del grupo, se impondrá la pena en su mitad superior, que podrá elevarse hasta la superior en grado.

5. En todos los casos, se impondrá además la pena de inhabilitación especial para profesión u oficio educativos, en el ámbito docente, deportivo y de tiempo libre, por un tiempo superior entre tres y diez años al de la duración de la pena de privación de libertad impuesta en su caso en la sentencia, atendiendo proporcionalmente a la gravedad del delito, el número de los cometidos y a las circunstancias que concurran en el delincuente.

6. El juez o tribunal acordará la destrucción, borrado o inutilización de los libros, archivos, documentos, artículos y cualquier clase de soporte objeto del delito a que se refieren los apartados anteriores o por medio de los cuales se hubiera cometido. Cuando el delito se hubiera cometido a través de tecnologías de la información y la comunicación, se acordará la retirada de los contenidos.

En los casos en los que, a través de un portal de acceso a internet o servicio de la sociedad de la información, se difundan exclusiva o preponderantemente los contenidos a que se refiere el apartado anterior, se ordenará el bloqueo del acceso o la interrupción de la prestación del mismo.

Última actualización, publicada el 31/03/2015, en vigor a partir del 01/07/2015.

ARTÍCULO 22.4º

Son circunstancias agravantes:

Cometer el delito por motivos racistas, antisemitas u otra clase de discriminación referente a la ideología, religión o creencias de la víctima, la etnia, raza o nación a la que pertenezca, su sexo, orientación o identidad sexual, razones de género, la enfermedad que padezca o su discapacidad.

Última actualización, publicada el 31/03/2015, en vigor a partir del 01/07/2015.

ANEXO B

RESPONSABILIDAD DE LOS PRESTADORES DE SERVICIOS DE ALOJAMIENTO O ALMACENAMIENTO DE DATOS

1. Los prestadores de un servicio de intermediación consistente en albergar datos proporcionados por el destinatario de este servicio no serán responsables por la información almacenada a petición del destinatario, siempre que:

    a) No tengan conocimiento efectivo de que la actividad o la información almacenada es ilícita o de que lesiona bienes o derechos de un tercero susceptibles de indemnización, o

    b) Si lo tienen, actúen con diligencia para retirar los datos o hacer imposible el acceso a ellos.

2. La exención de responsabilidad establecida en el apartado 1 no operará en el supuesto de que el destinatario del servicio actúe bajo la dirección, autoridad o control de su prestador.

ANEXO C

DATOS DE LA OBTENCIÓN DE MUESTRAS DE DELITOS DE ODIO CONTRA LA POBLACIÓN ÁRABE Y/O MUSULMANA.

En este anexo se muestra un enlace al *bucket* S3 público en el que consultar los mensajes de odio obtenidos del experimento realizado con alumnos de la facultad de Derecho y Ciencias Sociales de la Universidad de Castilla-La Mancha en el campus de ciudad real.

- https://s3-eu-west-1.amazonaws.com/shade-tfg-test/hate_crimes.txt

El archivo también se encuentra en la carpeta del proyecto bajo el nombre “hate_crimes.txt”.

ANEXO D

CONVENIO EUROPEO DE DERECHOS HUMANOS: ARTÍCULOS DE INTERÉS

ARTÍCULO 10.1: Libertad de Expresión

Toda persona tiene derecho a la libertad de expresión. Este derecho comprende la libertad de opinión y la libertad de recibir o de comunicar informaciones o ideas sin que pueda haber injerencia de autoridades públicas y sin consideración de fronteras. El presente artículo no impide que los Estados sometan a las empresas de radiodifusión, de cinematografía o de televisión a un régimen de autorización previa.

ARTÍCULO 17: Prohibición del abuso de derecho

Ninguna de las disposiciones del presente Convenio podrá ser interpretada en el sentido de implicar para un Estado, grupo o individuo, un derecho cualquiera a dedicarse a una actividad o a realizar un acto tendente a la destrucción de los derechos o libertades reconocidos en el presente Convenio o a limitaciones más amplias de estos derechos o libertades que las previstas en el mismo.

ANEXO E

FACTORES DE POLARIZACIÓN PARA LA IDENTIFICACIÓN DE DELITOS DE ODIO

- La percepción de la víctima. Siguiendo las recomendaciones de la Comisión Europea contra el Racismo y la Intolerancia del Consejo de Europa (ECRI), la sola percepción o sentimiento, por parte de la víctima, de que el motivo del delito sufrido pueda ser racista, xenófobo o discriminatorio debe obligar a las autoridades a llevar una investigación eficaz y completa para confirmar o descartar dicha naturaleza. Esa percepción subjetiva de la víctima, no significa que finalmente el hecho deba calificarse de racista, xenófobo o discriminatorio, pero obliga a la policía judicial o a los fiscales o a los jueces de instrucción a su investigación. En este sentido, se expresa el Tribunal Europeo de Derechos Humanos en sentencias de fechas de 4 de marzo de 2008, de 31 de marzo de 2010, de 4 de marzo de 2011 y de 20 de octubre de 2015.

- La pertenencia de la víctima a un colectivo o grupo minoritario por motivos étnicos, raciales, religiosos, de orientación o identidad sexual etc.

- Discriminación y odio por asociación. La víctima puede no pertenecer o ser miembro del grupo objetivo, pero puede ser un activista que actúa en solidaridad con el colectivo. Igualmente, puede darse el caso de que la víctima se hallase en compañía de algunos de los miembros del grupo vulnerable. En definitiva, se trata de víctimas que sin pertenecer a un colectivo minoritario son deliberadamente escogidas por su relación con el mismo. Piénsese en hechos cometidos contra las parejas interraciales o grupos de amigos de diferentes orígenes nacionales, religiosos o étnicos o contra los miembros de una ONG que defienden los derechos de minorías.

- Las expresiones o comentarios racistas, xenófobos u homófobos, o cualquier otro comentario vejatorio contra cualquier persona o colectivo, por su ideología, situación de exclusión social, orientación religiosa, por ser persona con discapacidad, etc., que profiera el autor/es al cometer los hechos. En este caso, se recomienda que sean recogidas con toda su literalidad en las declaraciones de la víctima o los testigos.

- Los tatuajes, el vestuario o la estética del autor de los hechos. En muchos casos, estos elementos tendrán una simbología relacionada con el odio, y ayudarán acreditar y describir de forma gráfica el perfil del autor y la motivación del delito. En este sentido, las Fuerzas y Cuerpos de Seguridad deberán aportar informes fotográficos incorporados a los atestados reflejando todos estos datos.

- La propaganda, estandartes, banderas, pancartas, etc. de carácter extremista o radical que pueda portar el autor de los hechos o que puedan encontrarse en su domicilio. En este último supuesto, si se lleva a cabo un registro domiciliario. Todos estos efectos serán filmados o fotografiados para su incorporación al atestado.

- Los antecedentes policiales del sospechoso. Antecedentes que pueden derivarse por haber participado en hechos similares, o por haber sido identificado anteriormente por asistir a conciertos de carácter neo-nazi, de música RAC/OI, conferencias, reuniones o manifestaciones de carácter ultra caracterizadas por su hostilidad a colectivos minoritarios.

  La ley equipara los antecedentes penales españoles a los correspondiente a las condenas firmes de jueces o tribunales impuestas en otros Estados de la Unión Europea tendrán el mismo valor que las impuestas por los jueces o tribunales españoles salvo que sus antecedentes hubieran sido cancelados, o pudieran serlo con arreglo al Derecho español, a los efectos de la concurrencia de la agravante de reincidencia.

- Que el incidente haya ocurrido cerca de un lugar de culto, un cementerio o un establecimiento de un grupo considerado minoritario en la vecindad, como por ejemplo una asociación de defensa de derechos humanos u ONG.

- La relación del sospechoso con grupos ultras del fútbol. En este sentido, habrá que cruzar los datos con los que dispongan los coordinadores de seguridad de estadios de fútbol, y que se recogen en el Registro Central de Sanciones en materia de violencia, racismo, xenofobia e intolerancia en el deporte.

- La relación del sospechoso con grupos o asociaciones caracterizadas por su odio, animadversión u hostilidad contra colectivos de inmigrantes, musulmanes, judíos, homosexuales, etc.

- La aparente gratuidad de los actos violentos, sin otro motivo manifiesto. Este factor debe ser considerado como un indicio muy poderoso.

- Enemistad histórica entre los miembros del grupo de la víctima y del presunto culpable.

- Cuando los hechos ocurran con motivo u ocasión de una fecha significativa para la comunidad o colectivo de destino. Ejemplos a citar serían: un viernes, día de la oración para musulmanes, o un sábado para los judíos, el día del orgullo gay, etc.

- Cuando los hechos ocurran en un día, hora o lugar en el que se conmemora un acontecimiento o constituye un símbolo para el delincuente, como por ejemplo el 20 de abril, día del cumpleaños de Hitler.

- La conducta del infractor. Los infractores de delitos de odio, frecuentemente, suelen mostrar sus prejuicios antes, durante y después de la comisión de incidente discriminatorio.

  En ocasiones, los autores filman con sus teléfonos móviles los hechos y los cuelgan en Internet para jactarse de su acción o presumir ante sus amigos. En este sentido, será muy interesante el análisis de su teléfono móvil u ordenadores, previa autorización judicial, para obtener pruebas. Existen ejemplos de casos en que dichas grabaciones han demostrado ser importantes para establecer el motivo, facilitando información importante que permite a los investigadores reunir las pruebas que conducen a una

condena. Si bien, estas medidas no serán apropiadas en todos los supuestos, dependerá de la gravedad del delito.

ANEXO F

RABAT PLAN OF ACTION ON THE PROHIBITION OF ADVOCACY OF NATIONAL, RACIAL OR RELIGIOUS HATRED THAT CONSTITUTES INCITEMENT TO DISCRIMINATION, HOSTILITY OR VIOLENCE

(a) **Context**: Context is of great importance when assessing whether particular statements are likely to incite discrimination, hostility or violence against the target group, and it may have a direct bearing on both intent and/or causation. Analysis of the context should place the speech act within the social and political context prevalent at the time the speech was made and disseminated;

(b) **Speaker**: The speaker's position or status in the society should be considered, specifically the individual's or organization's standing in the context of the audience to whom the speech is directed;

(c) **Intent**: Article 20 of the International Covenant on Civil and Political Rights anticipates intent. Negligence and recklessness are not sufficient for an act to be an offence under article 20 of the Covenant, as this article provides for "advocacy" and "incitement" rather than the mere distribution or circulation of material. In this regard, it requires the activation of a triangular relationship between the object and subject of the speech act as well as the audience.

(d) **Content and form**:The content of the speech constitutes one of the key foci of the court's deliberations and is a critical element of incitement. Content analysis may include the degree to which the speech was provocative and direct, as well as the form, style, nature of arguments deployed in the speech or the balance struck between arguments deployed;

(e) **Extent of the speech act**: Extent includes such elements as the reach of the speech act, its public nature, its magnitude and size of its audience. Other elements to consider include whether the speech is public, what means of dissemination are used, for example by a single leaflet or broadcast in the mainstream media or via the Internet, the

frequency, the quantity and the extent of the communications, whether the audience had the means to act on the incitement, whether the statement (or work) is circulated in a restricted environment or widely accessible to the general public;

(f) **Likelihood, including imminence**: Incitement, by definition, is an inchoate crime. The action advocated through incitement speech does not have to be committed for said speech to amount to a crime. Nevertheless, some degree of risk of harm must be identified. It means that the courts will have to determine that there was a reasonable probability that the speech would succeed in inciting actual action against the target group, recognizing that such causation should be rather direct.

ANEXO G

EXPERIMENTO PARA LA OBTENCIÓN DE MUESTRAS DE DELITOS DE ODIO CONTRA LA POBLACIÓN ÁRABE Y/O MUSULMANA.

En el marco de un Trabajo de Fin de Grado desarrollado por la Escuela de Informática de Ciudad Real y el Instituto de Derecho Penal Europeo e Internacional, se pretende desarrollar un programa informático con el fin de detectar delitos de odio en las redes sociales.

De entre todos los delitos de odio, contenidos en el art. 510 del CP, nos interesan dos modalidades:

*El art. 501.1 a) del CP sanciona con penas de prisión de una a cuatro años y multa de seis a doce meses a: "Quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, la hostilidad, discriminación o violencia, contra un grupo, una parte del mismo o contra una persona determinada por razón de su pertenencia a aquél, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual por razones de género, enfermedad o discapacidad". * El art. 501 2 a) sanciona con la pena de prisión de seis meses a dos años y multa de seis a doce meses a: "Quienes lesionen la dignidad de las personas mediante acciones que entrañen humillación menosprecio o descrédito de alguno de los grupos a que se refiere el apartado anterior, o de una parte de los mismos, o de cualquier persona determinada por razón de su pertenencia a ellos …"

De entre los diversos colectivos o grupos sociales que pueden resultar amparados por el art. 510 del CP la herramienta informática identificará mensajes de odio que afecten al colectivo musulmán y/o árabe[1]. Existen dos razones que generan actualmente un gran peligro de que este colectivo sea objeto de delitos de odio. En primer lugar, la inmigración ilegal, una parte significativa de los ciudadanos extranjeros que se encuentran en nuestro país, ya sea con residencia legal o no legal, pertenecen a este colectivo. El rechazo que en determinados sectores de la población genera la inmigración se centra en mensajes contra este colectivo. En segundo lugar, y obviamente por razones de terrorismo. No es infrecuente que en el discurso de odio anti-islam se mezclen ambos argumentos. Para el diseño del

[1]En el experimento se habla de árabes, musulmanes e islamistas de manera indistinta por la naturaleza del mismo, no se pretende aunar en el mismo grupo a las personas que se sientan identificadas con uno o más de estos calificativos.

programa resulta imprescindible contar con un gran número de este tipo de mensajes en los que se recoja la variedad de estructuras lingüísticas y de expresiones que pueden cobijar mensajes de odio contra el islam o en general la población musulmana que se encuentra en nuestro país. Con esta finalidad exclusivamente académica y con el objeto además de crear una herramienta preventiva de esta clase de delitos solicitamos tu colaboración. Nos gustaría que en relación a cada uno de los supuestos planteados escribieras un o más mensajes de no más de cinco líneas expresando una opinión que pudiera ser calificada de incitación al odio, la violencia o la discriminación contra el colectivo árabe y/o musulmán. El lenguaje empleado puede ser desde luego soez y maleducado ("vamos a darle una paliza a esos moros de mierda", "hay que echar a los putos islamistas"), pero cuida también de que aproximadamente la mitad de tus ejemplos estén redactados en un lenguaje algo más convencional.

Escenarios:

- Escribe un comentario en el que animes a un número indeterminado de personas a ejercer actos de violencia contra el colectivo musulmán en general.
- Escribe un comentario en el que animes a un grupo concreto de personas a que realicen actos de violencia contra un grupo de musulmanes muy determinados (por ejemplo, una familia de musulmanes que viven en tú calle).
- Escribe un comentario en el que animes a un número indeterminado de personas a que discriminen al colectivo musulmán en un determinado espacio (por ejemplo, la sanidad, la educación, el empleo…).
- Escribe un comentario en el que sin mencionar palabras relacionadas con la violencia o la discriminación, menosprecies en general a las personas pertenecientes al colectivo musulmán o islamista, tachándolos a todos por ejemplo de terroristas o de personas que reciben excesivas ayudas por parte del Estado.
- Igual que en el apartado anterior pero en relación a un grupo determinado de musulmanes (por ejemplo, una familia de tu barrio …).
- Escribe una expresión que consideres que constituye un acto de menosprecio humillación o descrédito contra los musulmanes o islamistas en general, contra un grupo concreto o una persona perteneciente a ese grupo por razón de su pertenencia al mismo.

ANEXO H

CÓDIGO PARA LA CONFIGURACIÓN DEL ANALIZADOR DE FREELING.

```
def freeling_invokeConf(parser) :

    invokeConf = pyfreeling.invoke_options()

    #Selección del tipo de entrada
    invokeConf.InputLevel = pyfreeling.TEXT
    #Nivel de análisis
    invokeConf.OutputLevel = pyfreeling.DEP

    #Activación de los módulos de análisis morfológico
    #Modulo Map
    invokeConf.MACO_UserMap = False
    invokeConf.MACO_AffixAnalysis = True
    invokeConf.MACO_MultiwordsDetection = True
    invokeConf.MACO_NumbersDetection = True
    invokeConf.MACO_PunctuationDetection = True
    invokeConf.MACO_DatesDetection = True
    invokeConf.MACO_QuantitiesDetection  = True
    invokeConf.MACO_DictionarySearch = True
    invokeConf.MACO_ProbabilityAssignment = True
    invokeConf.MACO_CompoundAnalysis = False
    invokeConf.MACO_NERecognition = True
    invokeConf.MACO_RetokContractions = False
    invokeConf.NEC_NEClassification = True

    #Seleccion del algoritmo para la desambiguación
    invokeConf.SENSE_WSD_which = pyfreeling.UKB
    #Etiquetado gramatical de basado en el trigrama de Markovian
    invokeConf.TAGGER_which = pyfreeling.HMM

    #Parseador
    invokeConf.DEP_which = parser

    return invokeConf
```

```
def freeling_configuration(language, data_path):

    config = pyfreeling.config_options()
    #idioma con el que se va a procesar el texto
    config.Lang = language
    #path de los datos correspondientes al idioma a indicar
    ldata_path = data_path + "/share/freeling/" +
       config.Lang + "/"

    #Archivo de configuración para la tokenización
    config.TOK_TokenizerFile = ldata_path + "tokenizer.dat"
    #Archivo de configuracion para el splitter
    config.SPLIT_SplitterFile = ldata_path + "splitter.dat"
    #Configuración de las opciones para el analizador morfológico
    config.MACO_Decimal = ","
    config.MACO_Thousand = "."
    config.MACO_LocutionsFile = ldata_path + "locucions.dat"
    config.MACO_QuantitiesFile = ldata_path +
       "quantities.dat"
    config.MACO_AffixFile = ldata_path + "afixos.dat"
    config.MACO_ProbabilityFile = ldata_path +
       "probabilitats.dat"
    config.MACO_DictionaryFile = ldata_path + "dicc.src"
    config.MACO_NPDataFile = ldata_path + "np.dat"
    config.MACO_PunctuationFile = ldata_path +
       "../common/punct.dat"
    config.MACO_ProbabilityThreshold = 0.001

    #Archivo de configuración para la anotación de significados
    config.SENSE_ConfigFile = ldata_path + "senses.dat"
    #Archivo de configuración para la desambiguación de
    #significados mediante el algoritmo UKB
    config.UKB_ConfigFile = ldata_path + "ukb.dat"
    #Configuración de las opciones de etiquetado gramatical
    config.TAGGER_HMMFile = ldata_path + "tagger.dat"
    config.TAGGER_ForceSelect = pyfreeling.RETOK
    #Archivo de configuracion para el parseador de dependencias
    config.DEP_TreelerFile = ldata_path +
       "dep_treeler/dependences.dat"
    #Archivo de configuracion para el NEC
    config.NEC_NECFile = ldata_path +
       "nerc/nec/nec-ab-poor1.dat"
    #Configuración del parseador gramatical
    config.PARSER_GrammarFile = ldata_path +
       "chunker/grammar-chunk.dat"
    #Configuracion de parseador de dependencias
    config.DEP_TxalaFile = ldata_path +
       "dep_txala/dependences.dat"

    return config
```

ANEXO I

EVOLUCIÓN DE LA INTERFAZ DE USUARIO.

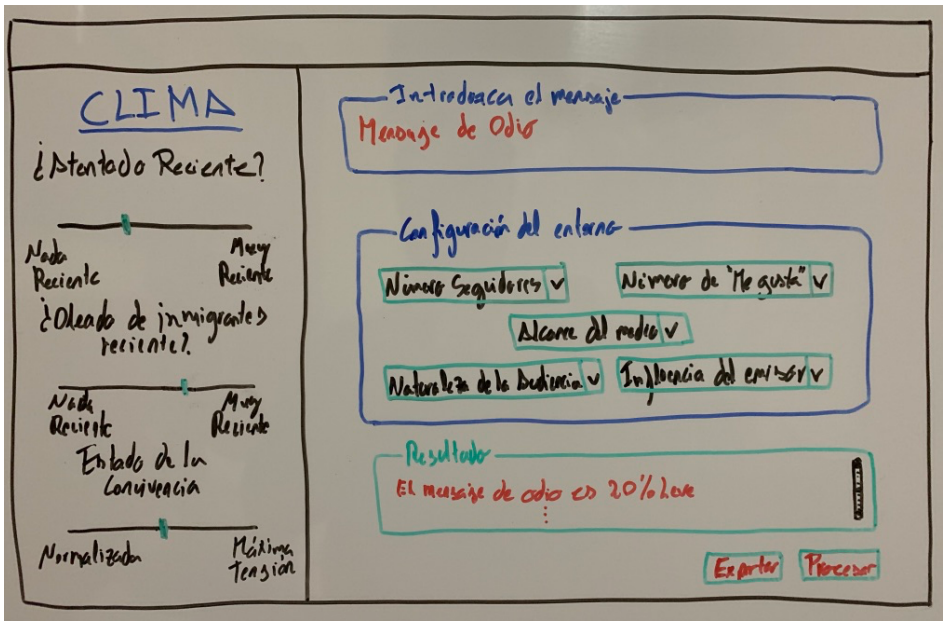


Figura I.1: Boceto inicial.

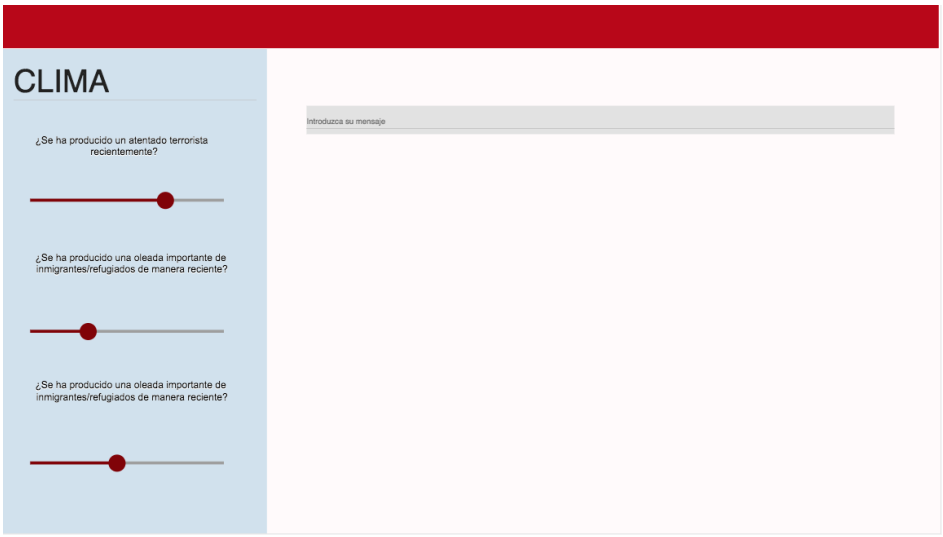


Figura I.2: Prueba de colores.

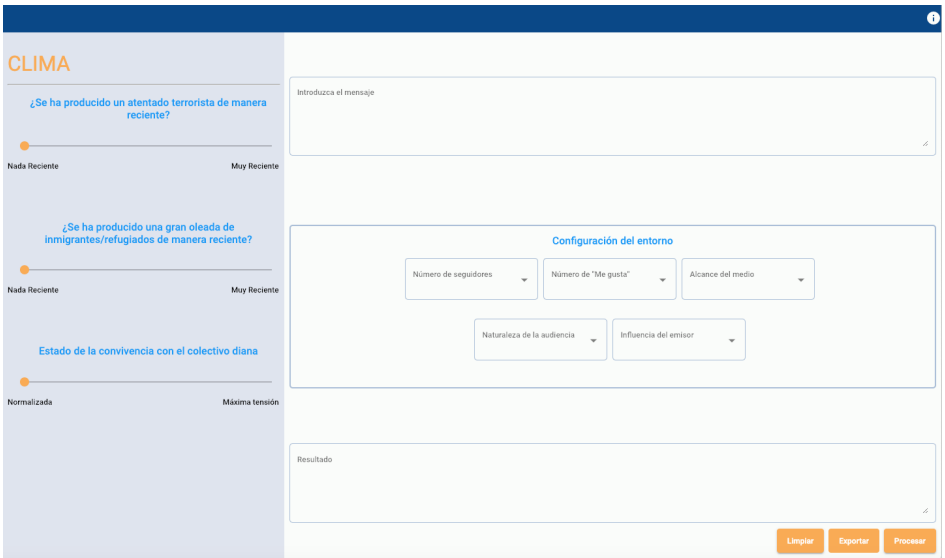


Figura I.3: Interfaz definitiva.

BIBLIOGRAFÍA

[1] Miguel Ángel Aguilar García. *Manual práctico para la investigación y enjuiciamiento de delitos de odio y discriminación*. 1st Edition. Generalitat de Catalunya: Centre d'Estudis Jurídics i Formació Especialitzada, 2015, págs. 1-397.

[2] Jordi Atserias, Elisabet Comelles y Aingeru Mayor. «TXALA un analizador libre de dependencias para el castellano». En: *Procesamiento del Lenguaje Natural* 35 (2005), págs. 455-456.

[3] Richard Bellman. *An introduction to artificial intelligence: Can computers think?* Boyd Fraser Publishing Company, 1978.

[4] David M Blei, Andrew Y Ng y Michael I Jordan. «Latent dirichlet allocation». En: *Journal of machine Learning research* 3.Jan (2003), págs. 993-1022.

[5] Margaret M. Bradley y Peter J. Lang. «Affective norms for English words (ANEW): Instruction manual and affective ratings». En: (1999).

[6] Thorsten Brants. «TnT: a statistical part-of-speech tagger». En: *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics. 2000, págs. 224-231.

[7] Leo Breiman. «Random forests». En: *Machine learning* 45.1 (2001), págs. 5-32.

[8] Peter F Brown y col. «Class-based n-gram models of natural language». En: *Computational linguistics* 18.4 (1992), págs. 467-479.

[9] Das deutsche Bundesgesetzblatt. *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken*. 2017. URL: https://goo.gl/Coqm8b (visitado 21-11-2018).

[10] Pete Burnap y Matthew L Williams. «Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making». En: *Policy & Internet* 7.2 (2015), págs. 223-242.

[11] Pete Burnap y Matthew L Williams. «Us and them: identifying cyber hate on Twitter across multiple protected characteristics». En: *EPJ Data Science* 5.1 (2016), pág. 11.

[12] Peter Burnap y Matthew Leighton Williams. «Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making». En: *Internet, Policy and Politics Conference, Oxford* (2014).

[13] Pete Burnap y col. «Detecting tension in online communities with computational Twitter analysis». En: *Technological Forecasting and Social Change* 95 (2015), págs. 96-108.

[14] Pete Burnap y col. «Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack». En: *Social Network Analysis and Mining* 4.1 (2014), pág. 206.

[15] David Caldevilla-Domínguez. «Impacto de las TIC y el 2.0: consecuencias para el sector de la comunicación». En: *Revista de Comunicación de la SEECI* 35 (2014), págs. 1-106.

[16] Miguel Ángel Cano Paños. «Internet y terrorismo islamista. Aspectos criminológicos y legales». En: (2008).

[17] Xavier Carreras. «Experiments with a higher-order projective dependency parser». En: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 2007, págs. 957-961.

[18] Jesús Bernal del Castillo. «La justificación y enaltecimiento del genocidio en la Reforma del Código Penal de 2015». En: *InDret* 2 (2016), págs. 1-22.

[19] Michael Chau y Jennifer Xu. «Mining communities and their relationships in blogs: A study of online hate groups». En: *International Journal of Human-Computer Studies* 65.1 (2007), págs. 57-70.

[20] Ying Chen y col. «Detecting offensive language in social media to protect adolescent online safety». En: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE. 2012, págs. 71-80.

[21] Gobinda G. Chowdhury. «Natural Language Processing». En: *Annual review of information science and technology* 37.1 (2003), págs. 51-89.

[22] Corinna Cortes y Vladimir Vapnik. «Support-vector networks». En: *Machine learning* 20.3 (1995), págs. 273-297.

[23] Maral Dadvar y col. «Improved cyberbullying detection using gender information». En: *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent. 2012, págs. 22-25.

[24] Maral Dadvar y col. «Improving cyberbullying detection with user context». En: *European Conference on Information Retrieval*. Springer. 2013, págs. 693-696.

[25] Tribunal Europeo de Derechos Humanos. *Convenio Europeo de Derechos Humanos*. 2010. URL: https://www.echr.coe.int/Documents/Convention_SPA.pdf (visitado 10-12-2018).

[26] Karthik Dinakar y col. «Common sense reasoning for detection, prevention, and mitigation of cyberbullying». En: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2.3 (2012), pág. 18.

[27] Nemanja Djuric y col. «Hate speech detection with comment embeddings». En: *Proceedings of the 24th international conference on world wide web* (2015), págs. 29-30.

[28] Ministerio de Empleo y Seguridad Social. *Recomendación de Política General Nº 15 relativa a la lucha contra el Discurso de Odio*. 2015. URL: http://www.mitramiss.gob.es/oberaxe/ficheros/documentos/2016_12_21-Recomendacion_ECRI_NO_15_Discurso_odio-ES.pdf (visitado 15-12-2018).

[29] Charniak Eugene y Drew McDermott. *Introduction to artificial intelligence*. 1985.

[30] Iginio Gagliardone y col. *Countering online hate speech*. Unesco Publishing, 2015.

[31] Alfonso Galán Muñoz. *Libertad de expresión y responsabilidad penal por contenidos ajenos en internet*. 1ª Edición. Tirant lo Blanch, 2010, págs. 1-277.

[32] Njagi Dennis Gitari y col. «A lexicon-based approach for hate speech detection». En: *International Journal of Multimedia and Ubiquitous Engineering* 10.4 (2015), págs. 215-230.

[33] John Haugeland. *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press, 1985.

[34] Minqing Hu y Bing Liu. «Mining and summarizing customer reviews». En: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), págs. 168-177.

[35] Akshay Java y col. «Why we twitter: understanding microblogging usage and communities». En: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM. 2007, págs. 56-65.

[36] Francisco P. Romero y Enrique Herrera-Viedma Jesús Serrano-Guerrero José Ángel Olivas. «Sentiment analysis: A review and comparative analysis of web services». En: *Information Sciences* 311 (2015), págs. 18-38.

[37] Ray Kurzweil. *The age of intelligent machines*. MIT press Cambridge, MA, 1990.

[38] Jon-Mirena Landa Gorostiza. *Los Delitos de Odio*. 1ª Edición. Tirant lo Blanch, 2018, págs. 1-152.

[39] Jon-Mirena Landa Gorostiza y col. *Delitos de Odio: Derecho Comparado y regulación Española*. 1ª Edición. Tirant lo Blanch, 2018, págs. 1-323.

[40] Patricia Laurenzo Copello. «La discriminación en el Código Penal de 1995». En: *Estudios Penales y Criminológicos, vol. XIX* (2012), págs. 221-288.

[41] Quoc Le y Tomas Mikolov. «Distributed representations of sentences and documents». En: *International Conference on Machine Learning*. 2014, págs. 1188-1196.

[42] Brian Levin. «From slavery to hate crime laws: The emergence of race and status-based protection in American criminal law.» En: *Journal of Social Issues, 58(2)* (2002), págs. 227-245.

[43] Bing Liu. «Sentiment analysis and opinion mining». En: *Synthesis lectures on human language technologies* 1 (2012), págs. 1-167.

[44] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. 1st Edition. Cambridge University Press, 2015, págs. 1-367.

[45] Xavier Lluís, Xavier Carreras y Lluís Màrquez. «Joint arc-factored parsing of syntactic and semantic dependencies». En: *Transactions of the Association for Computational Linguistics* 1 (2013), págs. 219-230.

[46] John McCarthy y col. «A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955». En: *Dartmouth College* (1955), pág. 12.

[47] Walaa Medhat, Ahmed Hassan y Hoda Korashy. «Sentiment analysis algorithms and applications: A survey». En: *Ain Shams Engineering Journal* 5.4 (2014), págs. 1093-1113.

[48] Yashar Mehdad y Joel Tetreault. «Do Characters Abuse More Than Words?» En: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2016, págs. 299-303.

[49] Tomas Mikolov y col. «Efficient estimation of word representations in vector space». En: *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)* (2013).

[50] Committee of Ministers to member States. *Recommendation No. R (97) 20*. 1997. URL: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b (visitado 12-12-2018).

[51] Fernando Miró Llinares. «El cibercrimen. Fenomenología y criminología de la delincuencia en el ciberespacio». En: *Marcial Pons, Madrid* (2012), págs. 1-332.

[52] Fernando Miró Llinares. «Taxonomía de la comunicación violenta y el discurso del odio en Internet». En: *Revista de Internet, derecho y política* 22 (2016), págs. 93-118.

[53] Fernando Miró-Llinares, Asier Moneva y Miriam Esteve. «Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments». En: *Crime Science* (2018), pág. 12.

[54] Fernando Molina Fernández. *Memento práctico*. 1ª Edición. Francis Lefebvre, 2017, págs. 1793-1814.

[55] Andrés Montoro y col. «An ANEW based Fuzzy Sentiment Analysis Model». En: *2018 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2018, Rio de Janeiro, Brazil, July 8-13, 2018*. 2018, págs. 1-7. DOI: 10.1109/FUZZ-IEEE.2018.8491492. URL: https://doi.org/10.1109/FUZZ-IEEE.2018.8491492.

[56] Nils Johan Nilsson. *Artificial intelligence: a new synthesis*. Morgan Kaufmann, 1998.

[57] Chikashi Nobata y col. «Abusive language detection in online user content». En: *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee. 2016, págs. 145-153.

[58] JAMES J NOLAN III, Yoshio Akiyama y Samuel Berhanu. «The Hate Crime Statistics Act of 1990: Developing a method for measuring the occurrence of hate violence». En: *American Behavioral Scientist* 46.1 (2002), págs. 136-153.

[59] David Lynton Poole, Alan K Mackworth y Randy Goebel. *Computational intelligence: a logical approach*. Vol. 1. Oxford University Press UK, 1998.

[60] Racism y European Commission against Intolerance. *ECRI General Policy Recommendation Nº. 15*. 2015. URL: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 (visitado 12-12-2018).

[61] Jonathan Rasmusson. *The Agile samurai: how agile masters deliver great software*. Pragmatic Bookshelf, 2010.

[62] Amir H Razavi y col. «Offensive language detection using multi-level classification». En: *Canadian Conference on Artificial Intelligence*. Springer. 2010, págs. 16-27.

[63] Elaine Rich y Kevin Knight. *Artificial intelligence*. 2nd Edition. McGraw-Hill, 1991.

[64] Samuel Rodríguez Ferrández. «El ámbito de aplicación del actual artículo 510 CP en retrospectiva y en prospectiva tras la reforma penal de 2015». En: *Revista de derecho penal y criminología* 12 (2014), págs. 165-232.

[65] Stuart Jonathan Russell y Peter Norvig. *Artificial intelligence: a modern approach*. 3ª Edición. Prentice hall Upper Saddle River, 2003, págs. 1-1132.

[66] Gabinete de Coordinación y Estudios. Secretaria de Estado de Seguridad. Ministerio del Interior. *Presentación del Informe 2016 sobre Incidentes relacionados con los Delitos de Odio en España*. 2016. URL: https://goo.gl/ZRpd2P (visitado 15-03-2018).

[67] Andrew Silke. «Holy warriors: Exploring the psychological processes of jihadi radicalization». En: *European journal of criminology* 5.1 (2008), págs. 99-123.

[68] Leandro Araújo Silva y col. «Analyzing the Targets of Hate in Online Social Media.» En: *Proceedings of the Tenth International Conference on Web and Social Media*. 2016, págs. 687-690.

[69] James Slagle y Michael R Wick. «A method for evaluating candidate expert system applications». En: *AI Magazine* 9.4 (1988), págs. 1-44.

[70] Sara Owsley Sood, Elizabeth F Churchill y Judd Antin. «Automatic identification of personal insults on social news sites». En: *Journal of the American Society for Information Science and Technology* 63.2 (2012), págs. 270-285.

[71] Ellen Spertus. «Smokey: Automatic recognition of hostile messages». En: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence AAAI'97/IAAI'97*. 1997, págs. 1058-1065.

[72] Anne Stenersen. «The Internet: A virtual training camp?» En: *Terrorism and Political Violence* 20.2 (2008), págs. 215-233.

[73] Mike Thelwall y col. «Sentiment strength detection in short informal text». En: *Journal of the American Society for Information Science and Technology* 61.12 (2010), págs. 2544-2558.

[74] Robin L Thompson. «Radicalization and the use of social media». En: *Journal of strategic security* 4.4 (2011), pág. 9.

[75] Cynthia Van Hee y col. «Detection and fine-grained classification of cyberbullying events». En: *International Conference Recent Advances in Natural Language Processing (RANLP)*. 2015, págs. 672-680.

[76] William Warner y Julia Hirschberg. «Detecting hate speech on the world wide web». En: *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics. 2012, págs. 19-26.

[77] Zeerak Waseem y Dirk Hovy. «Hateful symbols or hateful people? predictive features for hate speech detection on twitter». En: *Proceedings of the NAACL student research workshop*. 2016, págs. 88-93.

[78] Anne Weber. *Manual on hate speech*. 1st Edition. Council of Europe Publishing, 2009, págs. 1-96.

[79] Matthew L Williams y Pete Burnap. «Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data». En: *British Journal of Criminology* 56.2 (2015), págs. 211-238.

[80] Robert Andrew Wilson y Frank C Keil. *The MIT encyclopedia of the cognitive sciences*. MIT press, 1999.

[81] Patrick H Winston. *Artificial Intelligence*. Addison-Wesley, 1992.

[82] Guang Xiang y col. «Detecting offensive tweets via topical feature discovery over a large scale twitter corpus». En: *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM. 2012, págs. 1980-1984.

[83] Jun-Ming Xu y col. «Learning from bullying traces in social media». En: *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics. 2012, págs. 656-666.

[84] Lotfi A. Zadeh. «Fuzzy sets». En: *Information and control* 8.3 (1965), págs. 338-353.

ÍNDICE ALFABÉTICO