

PERFILADO DEL AUTOR Y VÍCTIMA DE DISCURSO DE ODIOS PARA ADAPTAR LA CONTRANARRATIVA A CADA TIPO DE PERSONA



Catálogo de publicaciones de la Administración General del Estado
<https://cpage.mpr.gob.es>



© Ministerio de Inclusión, Seguridad Social y Migraciones.
Madrid, 2023

Autores: El equipo de investigación y trabajo que se ha ocupado del desarrollo del proyecto dentro de IBiDat está formado por: Rosa Elvira Lillo, directora del equipo de investigación e investigadora principal (IP). Lara Quijano Sánchez, Irene Ramiro López, Iñaki Úcar, Patricia Callejo, José González Cabañas, miembros del equipo de investigación.

Edita y distribuye: Observatorio Español del Racismo y la Xenofobia
Calle Agustín de Betancourt, 11, séptima planta. 28003 Madrid
oberaxe@inclusion.gob.es
<https://www.inclusion.gob.es/oberaxe/es/index.htm>

NIPO 121-24-008-9

Diseño: Solana e Hijos, A.G., S.A.U.

Maquetación: Diseño Gráfico Gallego y Asociados, S. L.

La información y opiniones contenidas en este documento son responsabilidad de sus autores/as y no necesariamente reflejan la posición oficial del Ministerio de Inclusión, Seguridad Social y Migraciones.

ÍNDICE

Resumen	5
1. Introducción y motivación del proyecto	6
2. Vista general del estudio.....	7
3. Extracción de tweets.....	9
3.1. Necesidad de extraer nuevos datos	9
3.2. Problemas con Twitter	9
3.3. Método de extracción.....	10
3.4. Clasificación de los tweets.....	11
3.5. Extracción de los usuarios más influyentes y prolíficos	12
4. Estado del arte.....	14
4.1. Perfilado de usuarios	14
4.2. Grafo de usuarios.....	15
4.3. Análisis de los tweets.....	15
4.4. SocialHaterBert	15
5. Clasificación de los usuarios.....	16
5.1. Clasificación de usuarios.....	16
5.1.1. Análisis de las proporciones de odio y upstander	16
5.1.2. Análisis de conglomerados para clasificar a los usuarios	20
6. Clasificador de usuarios en función del perfil	28
6.1. Modelado por características del perfil.....	28
6.1.1. Detalles sobre algunas características del perfil.....	28
6.2. Clasificación con árbol de decisión.....	35
6.2.1. Funcionamiento de un Árbol de Decisión.....	35
6.2.2. Resultado	36



7. Estudio de los usuarios a partir de sus tweets	40
7.1. ¿Un usuario habla de temas diferentes cuando emite odio en comparación con cuando no lo hace?	40
7.1.1. ¿Un hater discrimina a un único colectivo o a varios?	43
7.1.2. Sobre los odiadores: ¿son sus tweets de odio más incendiarios que los que no son de odio?	46
7.1.3. ¿Los usuarios odiadores, neutros y upstander tienen diferentes personalidades?.....	47
8. Grafo de conversaciones	49
8.1. Diseño del grafo	49
8.2. Obtención de las aristas.....	50
8.3. Obtención de los nodos	51
8.4. Resultados	52
9. Resumen general de logros	54
9.1. Conclusiones	54
Anexo 1. Manual de uso del material entregado	57
Anexo 2. Sobre conversaciones en Twitter	68
Anexo 3. Grafo de personas conocidas.....	70
Bibliografía	72

Resumen

Este estudio se enmarca en el proyecto REAL-UP “Discurso de odio, racismo y xenofobia: mecanismos de alerta y respuesta coordinada”. (Hate speech, racism and xenophobia: Alert Mechanisms and Response, analysis of the Upstander speech) financiado por la Comisión Europea dentro del programa “Ciudadanos, Igualdad, Derechos y Valores” (CERV).

Está liderado por el Ministerio de Inclusión, Seguridad Social y Migraciones, a través del Observatorio Español del Racismo y la Xenofobia (OBERAXE), y tiene como socios al Ministerio del Interior, a través de la Oficina Nacional de Lucha contra los Delitos de Odio (ONDOD), la Comunidad de Investigación para la Excelencia (CREA) de la Universidad de Barcelona, el Grupo Internacional de Estudios sobre Teoría Crítica de la Universidad de Valencia, y CIDALIA, consultoría en diversidad. Y como asociada participa la asociación A-Digital, en la que se integran empresas prestadoras de servicios de alojamiento de datos como YouTube, Facebook, Instagram, Twitter o Microsoft.

En particular, el paquete de trabajo 3 (WP3) objeto de la investigación llevada a cabo por IBiDat (uc3m-Santander Big Data Institute), está liderado por la ONDOD. El objetivo general del trabajo es el estudio del perfil del autor de los mensajes de odio para adaptar la contranarrativa a cada tipo de persona, así como el de la víctima, para luchar de forma más efectiva contra el discurso de odio, previsto en el citado proyecto europeo.

1 Introducción y motivación del proyecto

Este estudio se enmarca en el proyecto REAL-UP “Discurso de odio, racismo y xenofobia: mecanismos de alerta y respuesta coordinada”¹ financiado por la Comisión Europea en el “Programa Derechos, Igualdad y Ciudadanía REC-RRAC-HATE-AG2017”.

El presente documento es la segunda y última entrega realizada por IBiDat y en él se describe el estado del arte sobre la utilización del perfil del autor y de los receptores de los mensajes ofensivos en los algoritmos de detección de discurso de odio y las estrategias para implementar este conocimiento u otro generado de nuevo. A lo largo de los siguientes capítulos, se introducirá al lector el análisis hecho hasta la fecha y el diseño de la solución planteada. En concreto, el [Capítulo 2](#) ofrece una vista general del proyecto. El [Capítulo 3](#) explica las necesidades, dificultades y alternativas sobre la extracción de nuevos tweets para llevar a cabo las tareas de este proyecto. El [Capítulo 4](#) detalla el estado del arte y trabajo previo sobre el que se basa este proyecto. Los capítulos siguientes engloban el trabajo elaborado en este proyecto: clasificación de los autores en Twitter ([Capítulo 5](#)), predicción del tipo de autor ([Capítulo 6](#)), perfilado de usuarios ([Capítulos 6 y 7](#)) y análisis de las interacciones entre usuarios ([Capítulo 8](#)). El [Anexo A](#) detalla el material generado y cómo utilizarlo. El [Anexo B](#) explica la dinámica de conversaciones en Twitter y qué parte de esto es importante para este proyecto. Finalmente, el [Anexo C](#) comenta el inicio de un análisis adicional que quizá interese explorar en el futuro.

1. <https://real-up.eu/>

2 Vista general del estudio

La metodología del estudio es la siguiente.

- **Clasificación de los Autores en Twitter** ([Capítulo 5](#))

El primer paso consistirá en definir los distintos tipos de autores, considerando la cantidad de mensajes de odio y de contranarrativa que cada uno escribe. Se clasificará a los usuarios según esta categorización para poder analizar los atributos específicos (cantidad de tweets, antigüedad en Twitter, temas de conversación, etc.) que caractericen a cada tipo de autor.

- **Predicción del Tipo de Autor dado los Datos del Perfil** ([Capítulo 6](#))

El segundo paso consistirá en poder predecir el tipo de autor conociendo únicamente la información de su perfil, como su nombre de usuario, el número de seguidores que tiene, a cuánta gente sigue, etc. El propósito es doble: no solo se pretende desarrollar una herramienta útil para el análisis requerido en este estudio, sino también determinar si la información contenida en el perfil de un usuario es suficiente para revelar si este emite una gran cantidad de mensajes de odio.

- **Perfilado de Usuarios** ([Capítulos 6 y 7](#))

Identificación de patrones y atributos del perfil y de los tweets que caractericen a cada autor, sobre todo, a los autores que emiten una gran cantidad de odio. Este proceso de perfilado proporcionará información valiosa para comprender mejor el comportamiento de los usuarios en la red social.

- **Grafo de Conversaciones** ([Capítulo 8](#))

Estudiar si el autor que emite una gran cantidad de odio escribe de manera individual o forma parte de un colectivo de odiadores con los que interactúa. En particular, qué tipo de colectivo; que dependerá de a qué comunidad discrimina (personas gitanas, judías, musulmanas, migrantes, etc.). Analizar si los usuarios que contestan a los mensajes ofensivos son también personas que vierten odio a la red, posibles víctimas o simplemente personas que interaccionan con los autores como meros observadores.

Para etiquetar los tweets que sean necesarios para llevar a cabo este análisis, se han usado algoritmos previamente desarrollados para el proyecto europeo REAL-UP, los cuales categorizan cada tweet según la [Tabla 1](#). Como nota final, a pesar de que en julio de 2023 la plataforma Twitter cambió su nombre a X, en este informe se continua haciendo referencia a ella utilizando su denominación original.

Tabla 1. Categorías del discurso de odio diseñadas para la clasificación de tweets dentro del proyecto.

ofensivo-odio	Discurso que representa ofensas personales o colectivas, incita a la discriminación reproduciendo tópicos o falsedades
extremo-odio	Discurso que defiende, incita y propone la ejecución de acciones violentas y hostiles contra los diferentes colectivos, de manera estereotipada y agresiva
neutro	Es un discurso descriptivo en el que no aparece odio, pero tampoco se posiciona en contra del acoso
upstander	Discurso alternativo que contribuye a una contranarrativa, rompiendo con los tópicos o posicionándose en defensa de las víctimas. También se denomina <i>counter speech</i>

3

Extracción de tweets

En esta sección se describe la necesidad de extraer información nueva de la red para poder llevar a cabo el análisis acordado. En segunda instancia, se informa de las dificultades para lograr esta extracción ocasionadas por la clausura del acceso gratis a la API de Twitter. Finalmente, se describen las alternativas usadas para sobrepasar este contratiempo.

3.1. Necesidad de extraer nuevos datos

La necesidad de extraer nuevos datos radica en dos razones fundamentales. La primera es que para analizar las interacciones entre usuarios en Twitter, es esencial conocer las respuestas a cada tweet que sea de interés. Y no solo las respuestas, sino también qué se contesta a éstas, y cómo se responde a continuación (lo que en Twitter se denomina conversación, ver el [Anexo B](#)). Esta dialéctica en línea nos permitirá conocer cómo interaccionan dos usuarios, si se suelen apoyar emitiendo odio o emitiendo contenido upstander, o si por el contrario, cualquier mensaje relacionado con el discurso de odio genera controversia en los comentarios. La segunda razón por la que resulta necesario extraer nuevos tweets es para obtener un registro completo (o todo lo completo posible) de los tweets de un autor para así poder perfilar a los usuarios en función de cuánto odio emiten en general y no en un momento particular. Lo que denominamos una imagen “estática” de la vida en la red social de un autor concreto. Recordamos que el objetivo del proyecto es dar seguimiento al proceso de descubrimiento de los patrones generales de comportamiento de diferentes usuarios en Twitter.

En conclusión, para llevar a cabo los objetivos de este proyecto, necesitamos extraer nueva información de Twitter a través de la API.

3.2. Problemas con Twitter

A fecha de 9 de Febrero de 2023 [\[1\]](#), sólo meses después de que Elon Musk adquiriera Twitter, el acceso gratuito a la API fue cesado. El único plan sin coste no permite extraer tweets, únicamente publicar, siendo su uso equivalente al de la interfaz de la web o de la aplicación de móvil. A modo informativo, la Tabla 2 muestra el coste de los distintos planes de acceso a la API actualmente.

Tabla 2: Coste de los planes de la API de Twitter a día 25/10/2023.

Plan	Coste por mes
Basic	100 USD
Pro	5000 USD
Enterprise	42000 USD (precio base)

Esta interfaz era el medio a través del cual todas las herramientas existentes accedían a la información de la red social. Con el cierre de dicha API, este proyecto se ha visto obligado a adaptar los objetivos acordados de la siguiente manera: el análisis de los usuarios y del entorno con el que interacciona se hará con una “foto estática” de Twitter en un periodo de tiempo determinado, es decir, extraeremos tweets de la red durante un periodo determinado y los analizaremos para obtener una descripción estadística de los usuarios y sus interacciones. Responderemos a las cuestiones planteadas sobre cómo varían las características de un usuario en función del odio que emite y las interacciones sociales entre emisores de odio y defensores de las minorías discriminadas, todo ello como resultado del análisis de estos tweets. También desarrollaremos herramientas que, dado la información de un usuario o de una red de usuarios, den como salida información relevante sobre los mismos. Sin embargo, no desarrollaremos herramientas dinámicas que, antes de ejecutar este análisis, extraigan directamente datos de Twitter, pues carecemos de acceso continuo a la API y, por lo tanto, no podríamos garantizar su correcta ejecución.

Afortunadamente, tuvimos una ventana desde julio hasta septiembre en la que la cuenta de la API de uno de los miembros del proyecto nos permitió realizar ciertas descargas. De cara a futuras extracciones, hacemos notar que desde Septiembre la única forma de acceder a los datos es con el plan de uso de la API anteriormente descrito. Sin embargo, obtuvimos datos suficientes para poder desarrollar el proyecto. Cabe destacar que esta licencia no estaba sujeta a las especificaciones de ningún plan vigente ni obsoleto. Por ejemplo, el límite era de 150 millones de tweets al mes, mucho más alto que el tope de 10000 y 1 millón de tweets de los planes Basic y Pro, respectivamente. Sin embargo, la búsqueda de tweets con antigüedad mayor a una semana estaba restringida. Esto implicó que no pudimos acceder a los tweets originales del proyecto ALRECO para obtener las respuestas a los mismos, por lo tanto, la solución que tomamos fue la de extraer nuevos tweets de la red, lo que hace a este proyecto novedoso y retador.

3.3. Método de extracción

Hemos obtenido tweets entre los días 23/07/2023 y 05/08/2023 (dos semanas) a partir de las palabras clave de odio de ALRECO. Para ello, fue necesaria la implementación de una metodología nueva de extracción, ya que la desarrollada en ALRECO era válida para la API v1.1, mientras que el acceso que teníamos se correspondía con la v2. Están clasificadas en los siguientes conjuntos: Globales, Minorías culturales y etnias religiosas, personas gitanas, personas judías, migrantes, personas musulmanas.

La consulta a la API de Twitter nos devuelve todos los tweets que no tienen más de una semana de antigüedad y que contienen una de estas palabras clave o *keywords*. No era posible filtrar por coordenadas. Lo mejor que pudimos hacer fue filtrar por *lang:es*, además de eliminar *retweets con -is:retweet*. En total, hemos extraído el siguiente número de tweets (Tabla 3).

Tabla 3: Número de tweets extraídos por categoría y cuántos de estos contienen respuestas que se puedan analizar.

Categoría	Número de tweets extraídos	Número de tweets con respuestas
Globales	396449	157810 (40%)
Minorías culturales y etnias religiosas	410774	187914 (46%)
Personas gitanas	16543	6649 (40%)
Personas judías	114496	32367 (28%)
Migrantes	616976	269941 (44%)
Personas musulmanas	116895	47171 (40%)

Nótese que las proporciones de tweets con respuestas se acercan al 40% en todos los casos excepto en el de personas judías. Tras leer múltiples tweets extraídos con keywords de este conjunto, se observa que muchos de ellos usan la palabra nazi para insultar a otros usuarios o personas notorias, en vez de tener relación con el discurso de odio contra o favor del colectivo judío. Podemos pensar que esta puede ser la razón por la que la proporción difiere de 40%.

3.4. Clasificación de los tweets

Una vez realizada la extracción se procede a ejecutar la clasificación de los tweets. Para poder realizar esta tarea es necesaria la creación de un *pipeline* que sirva para evaluar como entrada un tweet y poder clasificarlo según el modelo de HaterBERT preentrenado en su mejor resultado [2]. Se deja preseleccionado en el pipeline este modelo ya que es el que mejor resultado ofrece. Posteriormente se hace uso del *pipeline* definido y procedemos a la clasificación de nuestro conjunto de datos. En la Tabla 4 se muestran los resultados.

Se observa que la clase *extremo-odio* es la clase minoritaria con diferencia.

Tabla 4: Número de tweets extraídos agrupados por conjunto de keywords y clasificados en función del tipo de odio.

	Número de tweets por categoría			
	Odio extremo	Odio ofensivo	Upstander	Neutro
Globales	4 (0,001%)	120630 (30,4%)	11336 (2,9%)	264479 (66,7%)
Minorías culturales y etnias religiosas	6 (0,001%)	158717 (38,6%)	17608 (4,2%)	234443 (57,1%)
Personas gitanas	1 (0,006%)	8876 (53,7%)	1891 (11,4%)	5775 (34,9%)
Personas judías	11 (0,01%)	61536 (53,7%)	8785 (7,7%)	44164 (38,6%)
Migrantes	18 (0,003%)	213012 (34,5%)	35414 (5,7%)	368532 (59,7%)
Personas musulmanas	29 (0,02%)	52419 (44,8%)	5817 (5%)	58630 (50,2%)

3.5. Extracción de los usuarios más influyentes y prolíficos

Como se menciona en el [apartado 3.1.](#), también necesitamos una muestra de usuarios de los cuales conozcamos sus últimos tweets. En este capítulo describimos el proceso y los criterios seguidos para este fin.

Consideramos los autores de los tweets recopilados en el proceso de extracción mencionado anteriormente. El objetivo era identificar a los usuarios más influyentes o más prolíficos, pues suelen tener una presencia más sólida en la plataforma y, por lo tanto, pueden ofrecer una muestra de datos más completa y representativa². Primeramente, descartamos las cuentas que habían publicado menos de 30 tweets escritos desde su creación. En segundo lugar, excluimos a los que escriban con una frecuencia menor de 1 tweet por mes, es decir, aquellas cuyo ratio entre el número de tweets y los meses desde la creación de la cuenta fuera igual o inferior a uno. Este proceso nos aseguró que los usuarios que seleccionásemos a continuación habían estado en Twitter durante un período de tiempo significativo y que mantenían una actividad constante, tanto en términos absolutos como relativos.

A continuación, separamos la muestra de tweets en tipos de odio. Como la cantidad de tweets de odio extremo es considerablemente pequeña, los agrupamos junto con los de odio ofensivo. A partir de esto, calculamos el ranking de usuarios atendiendo a los siguientes criterios: Número de tweets de los cuales el usuario es autor (es decir, cuántas veces ha usado el usuario una palabra relacionada con el discurso de odio durante el periodo de extracción); Número de reacciones (suma de likes, retweets, respuestas y citas); Número de seguidores; Número de visualizaciones; Ratio del número de reacciones versus visualizaciones.

2. Nos estamos refiriendo a una identificación con el fin de dar seguimiento al proceso de descubrimiento de los patrones generales de comportamiento y no responde a una identificación personal que revele la identidad del autor del tweet o autores, respetando en todo momento el RGPD vigente.

De esta manera, generamos cinco rankings distintos, y seleccionamos los primeros 200 usuarios de cada uno, asegurándonos de no duplicar usuarios en el proceso. Como resultado, obtenemos tres listas diferentes: una con 1000 usuarios que han publicado al menos un mensaje de odio, otra con 1000 usuarios que han publicado un mensaje de apoyo (upstander), y una tercera con 1000 usuarios que han publicado al menos un mensaje neutro. A continuación, procedemos a extraer los tweets más recientes de cada uno de estos usuarios y con ello ya disponemos del dataset necesario para realizar los objetivos de este proyecto.



4 Estado del arte

Hasta este momento, la búsqueda en la literatura de artículos relacionados con el análisis de usuarios y su red social se ha centrado únicamente en la plataforma de Twitter. Existen otras redes sociales, como Reddit o Instagram, que tienen su propia interfaz con la que se pueden extraer datos de la red, pero hemos preferido acotar el estudio del estado del arte a la red social con la que trabajamos en este proyecto. De esta manera, estudiamos cómo otros artículos han usado las características propias de Twitter para alcanzar el objetivo que cada uno propone y conocemos los resultados que sabemos que se aplican para la red social que nos concierne.

4.1. Perfilado de usuarios

La literatura relativa al perfilado de usuarios en Twitter es variada; existen múltiples estudios que analizan las características de un perfil social para alcanzar diversos objetivos. Por ejemplo, el artículo *Understanding User Profiles on Social Media for Fake News Detection* [3] selecciona grupos representativos de usuarios propensos a creer en las noticias falsas y a continuación realiza un análisis comparativo de las características de los mismos. Observan que hay diferencias entre uno y otro grupo, y esto revela que las características de una cuenta son potencialmente útiles para distinguir un grupo de otro. En el proyecto que nos compete, llevar a cabo este proceso (separación por grupos y análisis de las características por grupo) es ciertamente interesante, pues si encontramos diferencias entre las mismas, podemos llegar a analizar las características que definen a los usuarios más odiadores y a los que no. Otra particularidad del artículo mencionado es que usa características “explícitas” e “implícitas”, las primeras siendo las que proporciona la API directamente (p. ej., número de seguidores y de *likes*) y las últimas siendo las que no (edad, género, personalidad). En nuestro caso, además de usar la información que viene dada, también procesaremos los datos para extraer otras características de interés.

Otro artículo relativo al perfilado de usuarios es *Early author profiling on Twitter using profile features with multi-resolution* [4]. En él, se pretende predecir características demográficas sobre los autores a partir de los tweets (por ejemplo, edad, sexo, lengua materna). En nuestro proyecto, también usaremos el registro de tweets de un usuario para extraer características que sean potencialmente interesantes para distinguir a los usuarios.

Nótese que los artículos citados en esta sección no tratan ninguno de modelar a los usuarios en función del odio que emiten ni tienen relación con el análisis de discurso de odio en Twitter. Hasta donde alcanza

nuestro conocimiento, este proyecto es pionero en el perfilado de usuarios en función del odio definido como está en este proyecto.

4.2. Grafo de usuarios

Para crear el grafo de usuarios existe una herramienta tecnológica de referencia, Graphext. Permite identificar las comunidades en una red mediante el análisis de *clusters*. Sin embargo, no es lo que precisamos para este trabajo, pues no permite pintar las aristas y los nodos de tal manera que representen el tipo de interacción y el tipo de usuario, respectivamente.

Una alternativa más clásica es el programa de visualización de grafos **Visone**. Podemos utilizarla para establecer múltiples conexiones entre los nodos (usuarios) y analizar sus interacciones. Podemos buscar patrones de comportamiento, como si los usuarios que emiten más odio tienden a comunicarse entre ellos o si interactúan principalmente con otros tipos de usuarios. Será esencial para comprender las dinámicas de conversación en la plataforma.

4.3. Análisis de los tweets

El grupo Cardiff NLP de la Universidad de Cardiff recientemente lanzó el Proyecto TweetNLP [5] [6]. TweetNLP es una plataforma centrada en el procesamiento de lenguaje natural (Natural Language Processing, NLP) y su aplicación en las redes sociales. Incluye múltiples tareas heterogéneas en Twitter, todas ellas enmarcadas en la clasificación multiclase de tweets. Entre ellas se incluye la clasificación de temáticas (seleccionadas cuidadosamente basándose en las tendencias de Twitter), análisis de sentimiento, detección de ironía, detección de odio, identificador de lenguaje ofensivo, predicción de *emojis* y reconocimiento de emoción.

La desventaja de estas herramientas es que la mayoría solo son válidas para tweets en inglés. Nuestro proyecto, por el contrario, solo contempla tweets escritos en lengua castellana. Si queremos usarlas, habremos de traducir los tweets.

4.4. SocialHaterBert

En el proyecto que precede a este estudio (mencionado en el [Capítulo 2](#)), se vio la importancia de la red social para categorizar los tweets de odio. El clasificador SocialHaterBERT [2] toma como entrada una gran cantidad de características inferidas de la cuenta de cada usuario y de la red social que le rodea. En particular, usaba atributos que actuaban como medidores de centralidad de un usuario (p. ej., *eigenvector*, el medidor de la influencia de un nodo en la red), atributos extraídos de los últimos 200 tweets del usuario (p. ej., *top_categories*, la lista de temáticas más frecuentes). Sobre esta base, seleccionaremos los atributos que nos interesen para el estudio actual además de generar más datos que sospechemos que tengan relación con el tipo de usuario.

5 Clasificación de los usuarios

El objetivo primero en esta sección es etiquetar a cada usuario en función de cuánto odio y cuánta contranarrativa emite, para así poder analizar si las demás características de un usuario dependen de los niveles de odio y contranarrativa. Además, construiremos un clasificador que determine de qué tipo es cada usuario, usando solamente la información del perfil.

5.1. Clasificación de usuarios

Esta sección trata de cómo se ha llegado a categorizar a los usuarios siguiendo técnicas de análisis de conglomerados (*cluster analysis*) para datos numéricos bivariados. *Cluster analysis* es una técnica estadística que agrupa datos similares en conjuntos o *clusters* con el fin de identificar patrones o estructuras dentro de un conjunto de datos. En nuestro caso, el objetivo de aplicar un algoritmo de *clustering* es obtener una clasificación de los usuarios en función de cuánto odio y cuánta contranarrativa emiten. Podemos imaginar, a priori, que un *cluster* se va a caracterizar por individuos muy odiadores y poco upstander. Otro *cluster* puede ser el opuesto, es decir, puede que contenga a usuarios muy upstander y poco odiadores. Quizá también se lleguen a distinguir dos grupos diferentes de odiadores: los que emitan odio de manera moderada y los que emitan odio con mucha frecuencia.

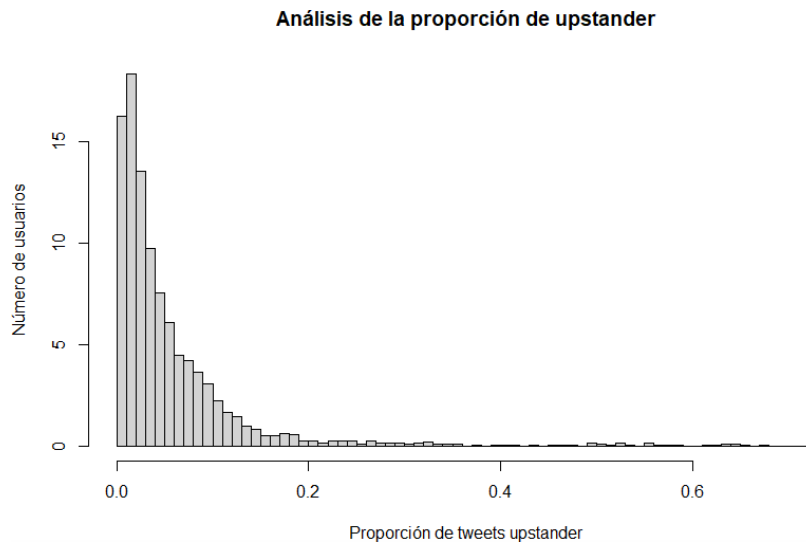
Partimos de un conjunto de 2722 usuarios, y para cada uno se ha extraído el porcentaje de mensajes de odio (tanto odio ofensivo como extremo) y el porcentaje de mensajes upstander. Como el porcentaje de neutros se infiere directamente de los anteriores, es una variable dependiente de las otras y por tanto no se ha considerado. Además, hemos agrupado odio ofensivo y extremo debido a que la cantidad de tweets de odio ofensivo es considerablemente pequeña ([ver Tabla 4](#)).

5.1.1. Análisis de las proporciones de odio y upstander

Los métodos de *clustering* frecuentemente asumen una estructura subyacente para los datos y buscan ajustar este modelo para identificar grupos distintos de entre sus datos. Es por ello que es importante analizar la distribución de los datos antes de aplicar un método de *clustering*, pues puede no seguir el modelo asumido y haya que compensarlo con esfuerzo adicional.

Primero comenzaremos analizando la distribución de las proporciones de tweets upstander. De ahora en adelante, esta variable será denotada por X .

Figura 1: Histograma de las proporciones de tweets clasificados como upstander dentro de la muestra de usuarios analizados



En la Figura 1 vemos el histograma de las 2722 muestras de X . A priori, es claro que no sigue una distribución normal, que es la más entendida y usada por los métodos de *clustering*. Para modelar proporciones, que están en el intervalo $[0, 1]$, se suele usar la distribución $\text{Beta}(\alpha, \beta)$ pues toma valores en ese intervalo. Por tanto, suponemos que $X \sim \text{Beta}(\alpha, \beta)$.

Ahora, sabiendo que:

$$E[X] = \frac{\alpha}{\alpha + \beta}$$

$$\text{Mediana}[X] \approx \frac{\alpha}{\alpha + \beta}$$

$$\text{Moda}[X] = \frac{\alpha}{\alpha + \beta}$$

podemos estimar los parámetros α y β usando los valores muestrales de la media, la mediana y la moda.

Figura 2: Histograma de las proporciones de tweets upstander acompañada de la función de densidad de una Beta($\alpha = 0,737585$, $\beta = 12,70973$), cuyos parámetros están inferidos a partir de la media y la mediana.

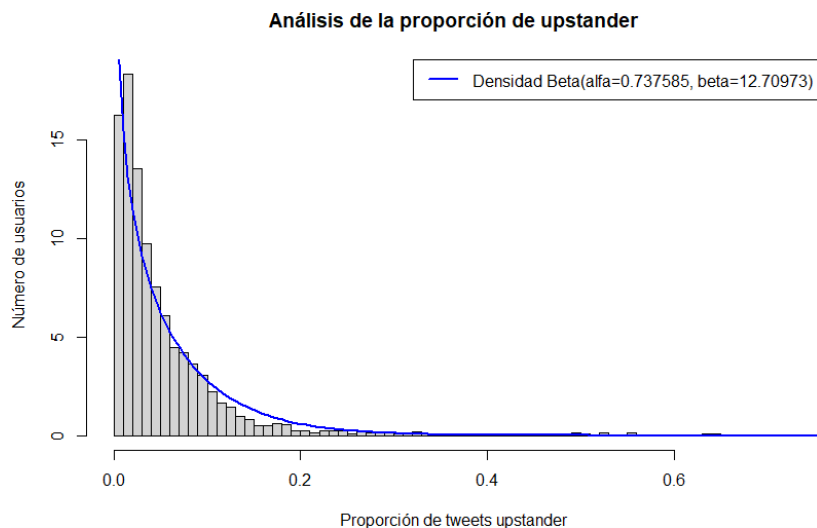
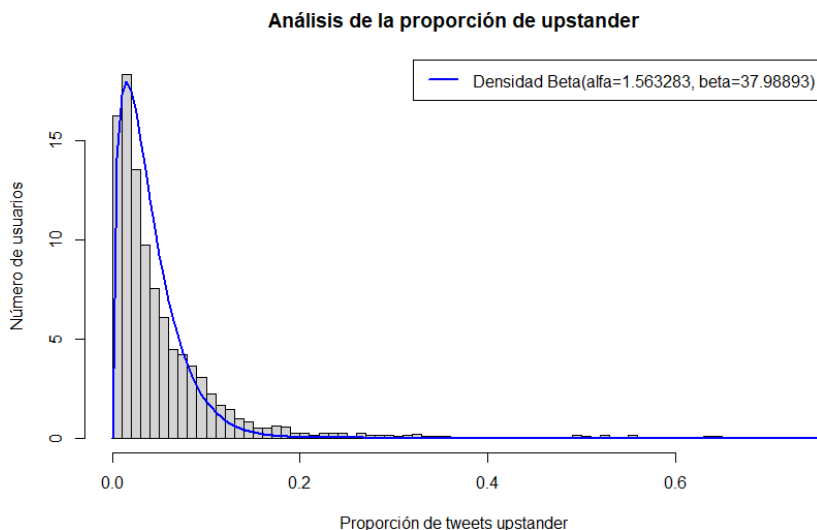
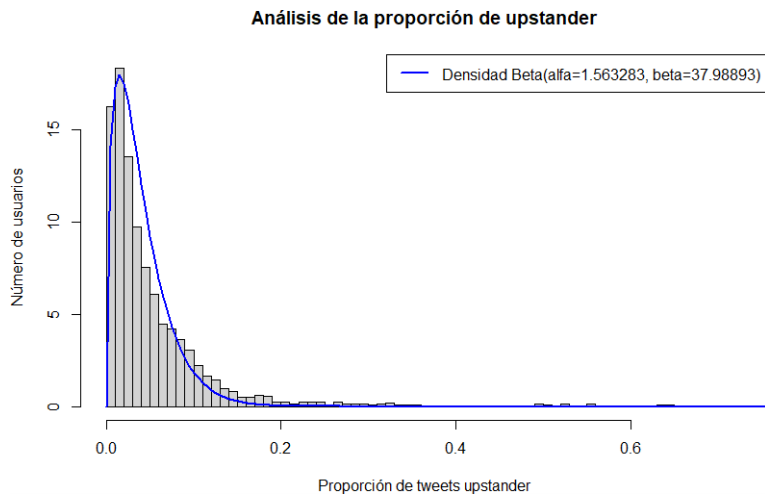


Figura 3: Histograma de las proporciones de tweets upstander acompañada de la función de densidad de una Beta($\alpha = 1,563283$, $\beta = 37,98893$), cuyos parámetros están inferidos a partir de la moda y la mediana.



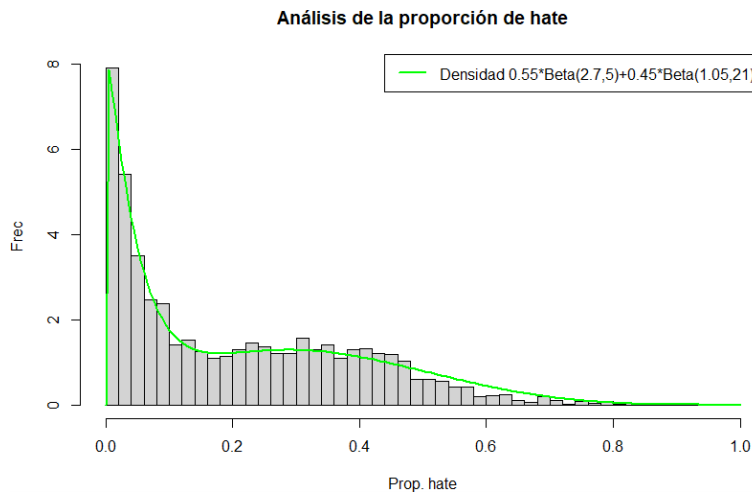
Se puede observar que tanto en la Figura 2 como en la Figura 3, la función de densidad se ajusta bastante bien al histograma. Sin embargo, en la Figura 2 se observa que la densidad converge más rápidamente a cero que el histograma. Al contrario ocurre en la Figura 3. Por tanto, podemos pensar que nuestra X es, en realidad, la suma de dos betas.

Figura 4: Histograma de las proporciones de tweets upstander acompañada de la función de densidad resultado de sumar las dos anteriores y ponderarlas por igual (con pesos de 0,5). En la leyenda, $\alpha_1 = 0,737585$, $\alpha_2 = 1,563283$, $\beta_1 = 12,70973$ y $\beta_2 = 37,98893$.



En la Figura 4 vemos que la densidad como suma de las betas anteriores (ponderadas por igual) se ajusta ahora mucho mejor al histograma tanto para valores cercanos al cero como para valores alejados del cero. Por tanto, es razonable asumir que X , la proporción de los mensajes upstander, se distribuye como la suma de dos betas. A continuación, analizamos la distribución de las proporciones de hate. De ahora en adelante, llamaremos a esta variable Y .

Figura 5: Histograma de las proporciones de tweets *hater* acompañada de la función de densidad $0,55 * \text{Beta}(\alpha = 2,7, \beta = 5) + 0,45 * \text{Beta}(\alpha = 1,05, \beta = 21)$.

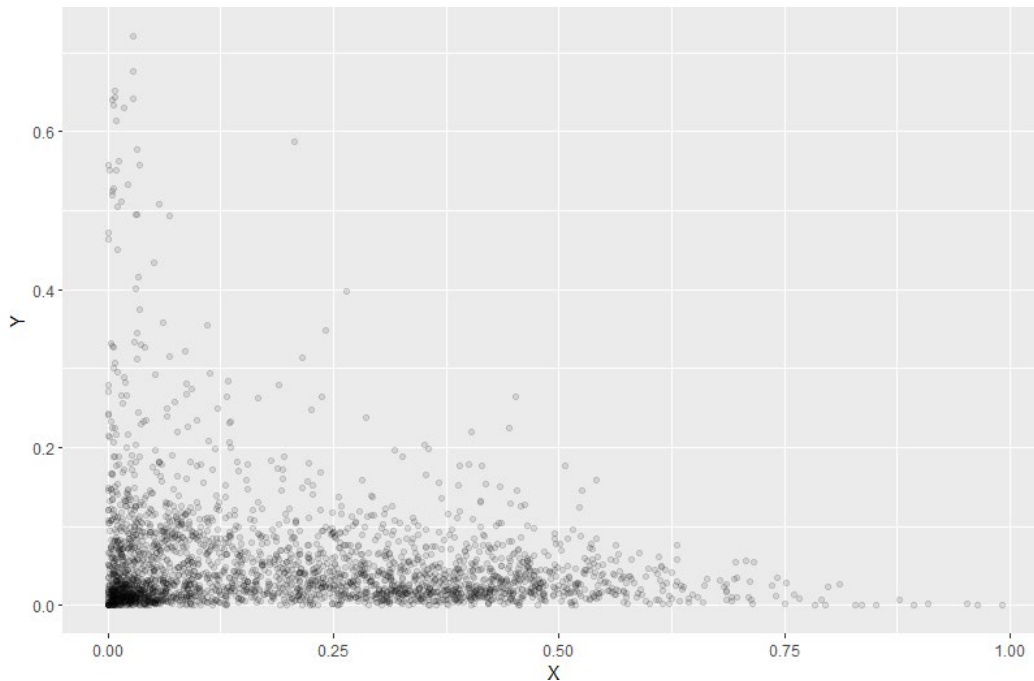


En la Figura 5 se muestra la densidad $0,55 * \text{Beta}(\alpha = 2,7, \beta = 5) + 0,45 * \text{Beta}(\alpha = 1,05, \beta = 21)$ sobre el histograma de las muestras de Y , estimado a mano. Vemos, de nuevo, que se ajusta bastante bien y por tanto es razonable suponer que Y , la proporción de los mensajes de odio, se distribuye como la suma de dos betas.

5.1.2. Análisis de conglomerados para clasificar a los usuarios

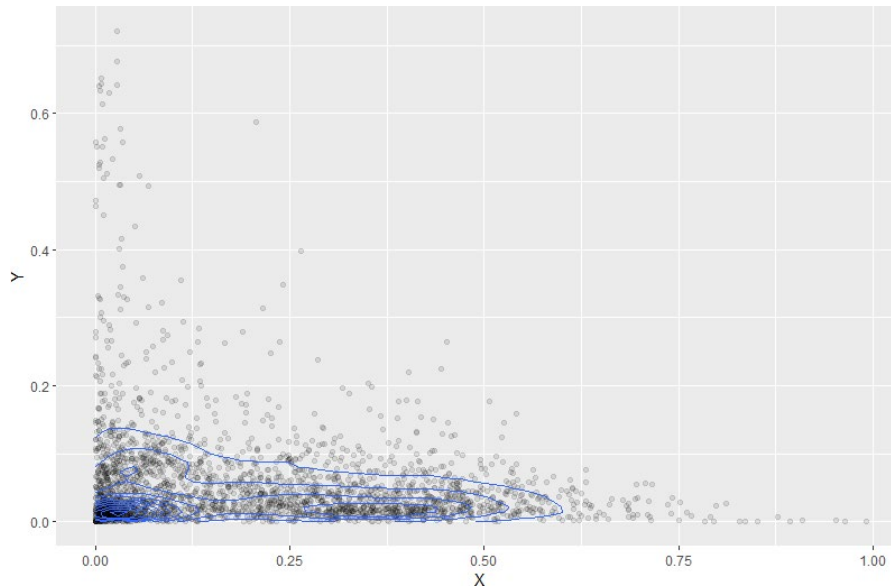
Veamos a continuación nuestros datos en el plano.

Figura 6: Gráfico de dispersión de los usuarios analizados contrastando la proporción de odio emitida frente a la de mensajes de tipo upstander.



La Figura 6 representa a cada usuario como un punto de coordenadas (x = proporción de odio, y = proporción upstander). Cuanto más alejado del eje X esté dispuesto un punto, más mensajes upstander escribe, y cuanto más alejado esté del eje Y, más odio emite. Se observa que ningún punto sobrepasa la diagonal $x + y = 1$ pues son proporciones que no pueden sumar más de uno. Los puntos han sido graficados no totalmente opacos sino con un nivel de transparencia para que nos permita ver en qué zonas hay mayor cúmulo de usuarios. Gracias a esto, se puede ver que cerca de la esquina inferior izquierda, donde los niveles de odio y upstander son valores cercanos al cero, se acumulan muchos puntos. Estos son los usuarios que publican pocos tweets relacionados con el discurso de odio: son los que más tweets neutros escriben.

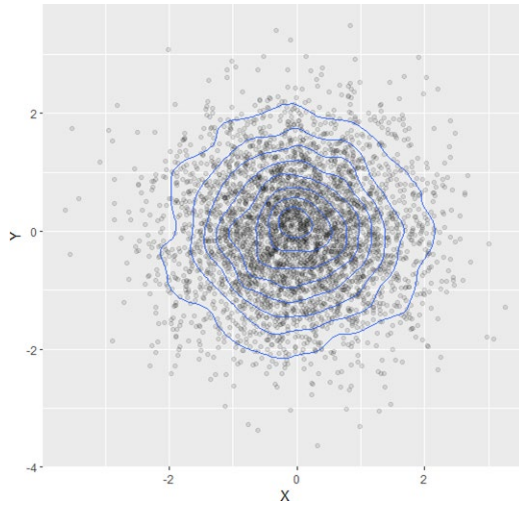
Figura 7: Estimación de la densidad para el gráfico de dispersión de los usuarios analizados contrastando la proporción de odio emitida frente a la de mensajes de tipo upstander.



La Figura 7 muestra el resultado de la estimación de la densidad del vector aleatorio (Y, X) con contornos. Este gráfico es similar a un mapa topográfico: las curvas de nivel representan cuánta densidad o “altura” hay en cada región, es decir, la concentración de usuarios. En particular, es muy evidente que cerca del $(0,0)$ aparece un “pico”, una curva cerrada y que no contiene a otras en su interior, lo que significa que en la región con baja proporción de odio y de upstander se concentran muchos usuarios. Esto ya se había observado en la [Figura 6](#). Pero ahora, se detectan otros dos “picos”: uno que está justo encima del mencionado anteriormente y otro alargado a la derecha. A priori, esto nos da indicios de que los datos se pueden clasificar en tres conglomerados.

Es importante recalcar que estos datos claramente no siguen una distribución normal bivariada, lo cual ya se intuía porque las proporciones de odio y upstander no son normales. En la [Figura 8](#) se muestra cómo sería la densidad estimada de una normal bivariada.

Figura 8: Estimación de la densidad para una normal bivariada simulada de manera aleatoria.



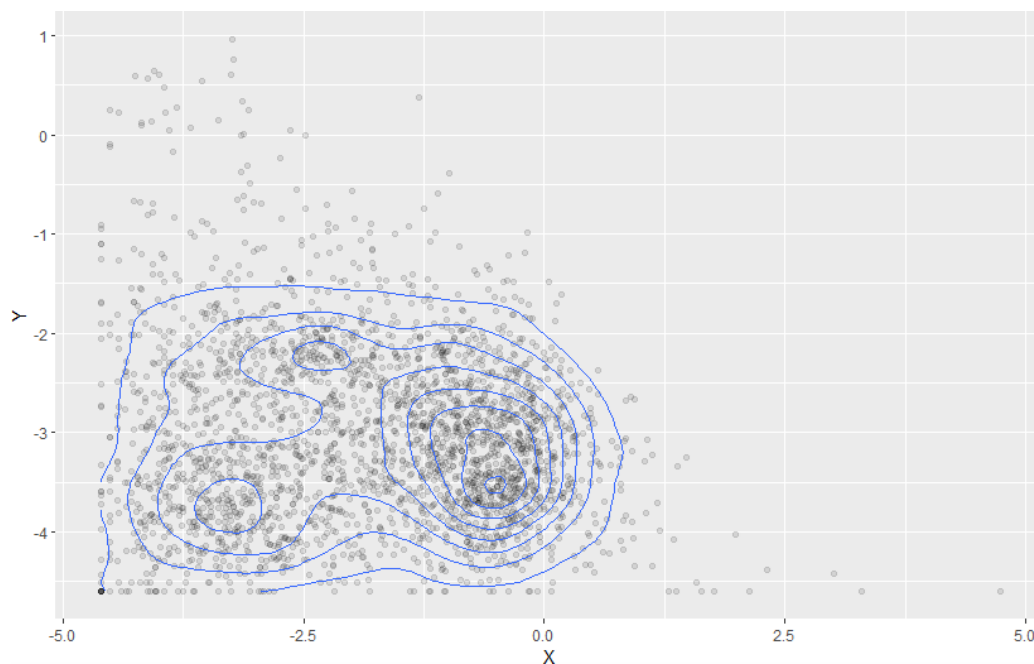
Esto dificulta la separación por *clusters* pues la mayoría de métodos de *clustering* asumen que los datos siguen una distribución normal multivariada. En particular, el algoritmo de K-medias, que se adapta bien a un gran número de muestras y se ha utilizado en una amplia gama de ámbitos de aplicación en muchos campos diferentes, funciona mejor cuando los datos se distribuyen en clusteres parecidos a la [Figura 7](#). Para sobrepasar este obstáculo, lo que hacemos es aplicar una transformación a nuestros datos.

La transformación elegida es la siguiente:

$$x \mapsto \frac{x}{1-x} \mapsto \log \frac{x}{1-x}$$

La primera transformación lleva el intervalo (0,1) a todo \mathbb{R}^+ , y la segunda lo lleva a todo \mathbb{R} , que es el conjunto donde una variable normal toma valores.

Figura 9: Gráfico de dispersión de las proporciones de odio emitido frente a la de mensajes de tipo upstander una vez realizadas las transformaciones descritas y acompañado de la estimación de la densidad.



En la Figura 9 vemos que, tras aplicar la transformación, los tres “picos” anteriores son más circulares, se parecen más a la [Figura 8](#), y por ende, más parecidos a una distribución normal. A priori, esto significa que están mejor predispuestos a que se cumplan las hipótesis del algoritmo de K-medias. Veamos cómo se comporta este algoritmo con los datos antes y después de transformar.

Figura 10: Gráfico de dispersión de las proporciones agrupadas en 3 clusters, habiéndolas agrupado sin transformar los datos.

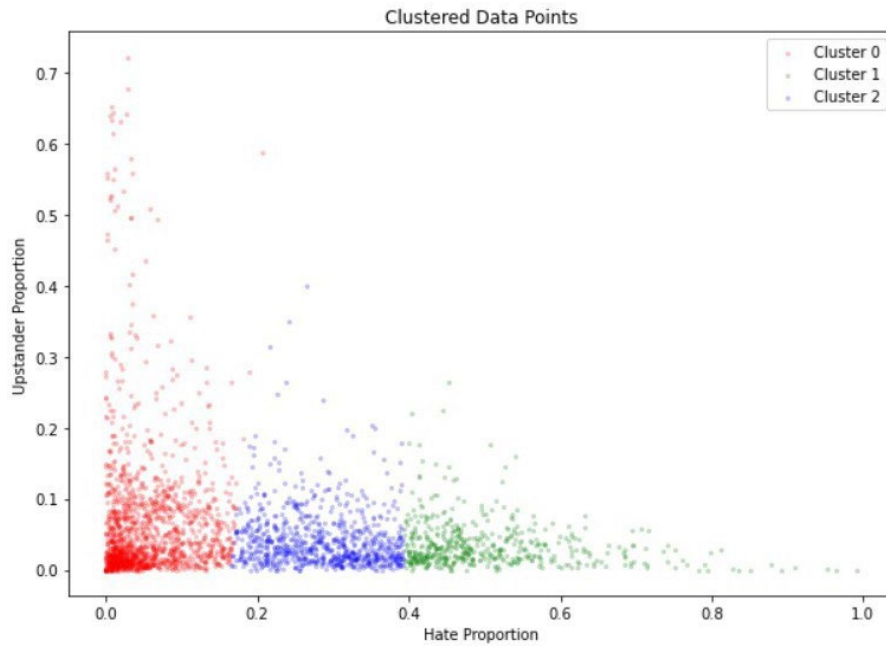
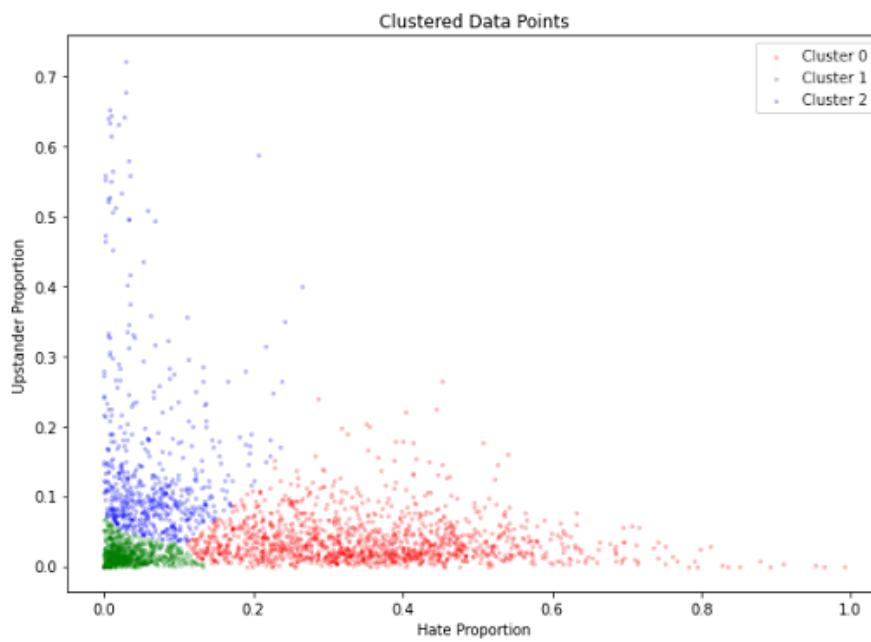


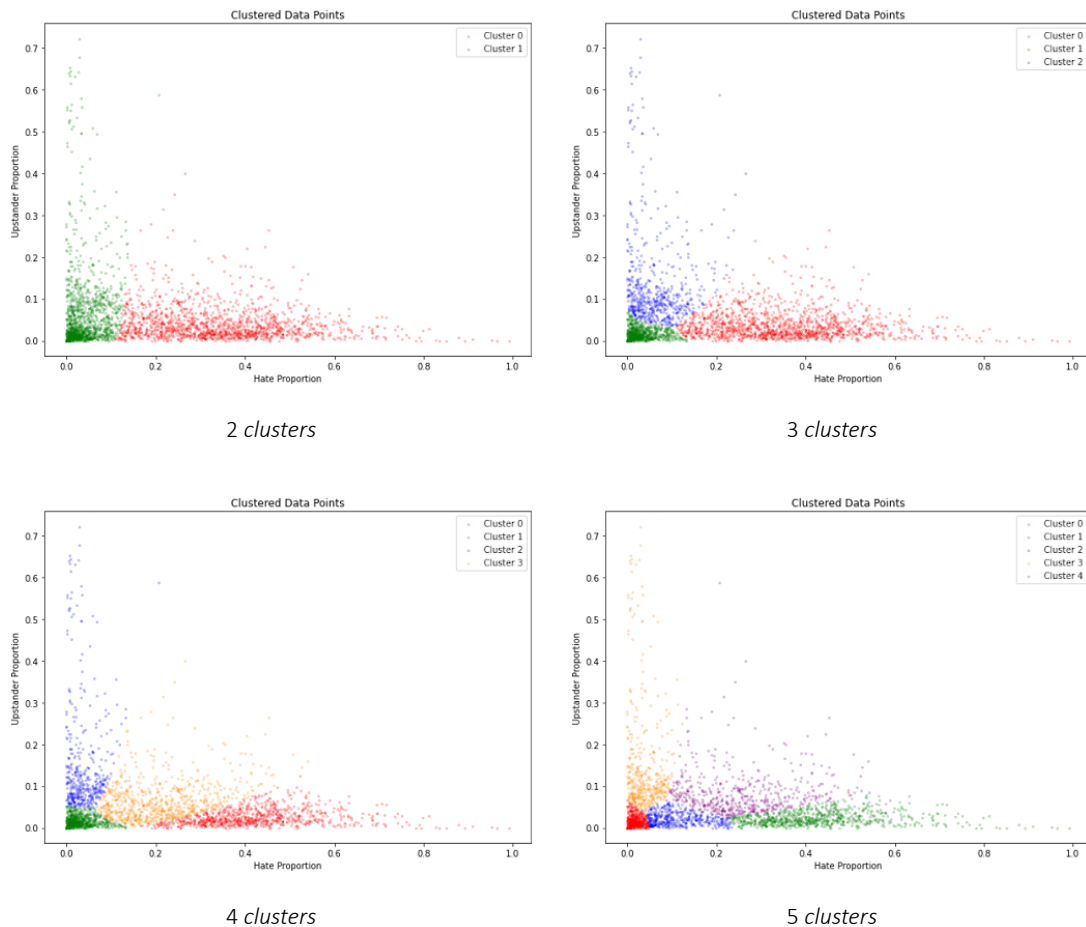
Figura 11: Gráfico de dispersión de las proporciones agrupadas en 3 clusters, habiéndolas agrupado con los datos transformados.



Observamos que los *clusters* en la [Figura 11](#) muestran una mayor similitud con los patrones previamente observados en la [Figura 7](#), puesto que se diferencian tres conglomerados: usuarios con poco nivel de odio y upstander (en verde), usuarios con mayor nivel de odio (en rojo) y usuarios con mayor nivel de mensajes upstander. En contraste, los conjuntos de datos en la [Figura 10](#) parecen diferenciarse principalmente en función del nivel de odio, sin hacer uso de las proporciones upstander para la separación. Si usáramos esta división, estaríamos etiquetando a los usuarios en función de si emiten poco, moderado o mucho odio. Esto no nos interesa si también deseamos caracterizar a los usuarios que actúan como “upstanders”. Por tanto, concluimos que aplicar la transformación antes de aplicar K-medias es el método que nos da resultados más fieles a la estructura subyacente de los datos, además de proporcionarnos unos grupos de usuarios que son más convenientes para nuestro análisis.

Pasamos ahora a la elección del número más indicado de *clusters*. En el análisis anterior habíamos especificado que este número fuera tres, pero podemos observar cómo se comporta el algoritmo de K-medias para un número distinto.

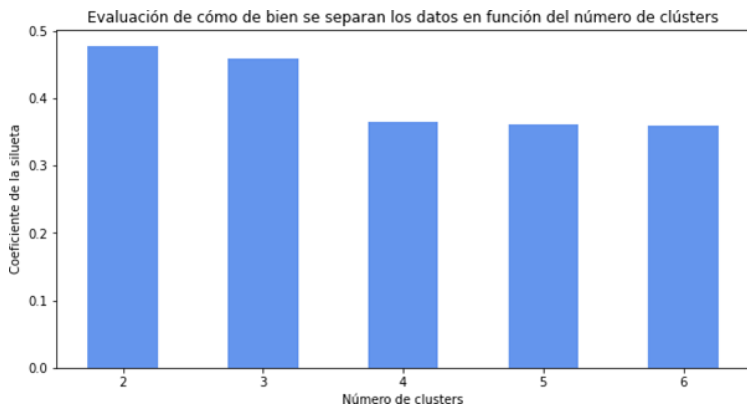
Figura 12: El algoritmo K-medias aplicado a los datos transformados para distinto número de *clusters*.



En la Figura 12 podemos ver el resultado de aplicar el algoritmo K-medias a los datos transformados para 2, 3, 4 y 5 *clusters*. En el caso de 2 *clusters*, vemos que la separación ha sido hecha prácticamente teniendo en cuenta únicamente el nivel de odio, puesto que separa a los que emiten aproximadamente más de 10% de odio (en rojo) de los que no (en verde). En el caso de 3 *clusters*, de nuevo, vemos que los usuarios se dividen en aquellos que tienen poca relación con el discurso de odio (en verde), aquellos que emiten más odio (en rojo) y aquellos que publican más contenido upstander (en azul). En el caso de 4 *clusters*, vemos que aparece un grupo en medio de entre estos dos últimos, el que emite tanto contenido odiador como upstander (en naranja). Por último, en el caso de 5 *clusters*, vemos que este grupo intermedio se ha dividido en función de si escribe más tweets upstander (en violeta) o menos (en azul).

A continuación, para computar cuál es el número de *clusters* que mejor se ajusta a los datos, es decir, que “mejor separa los datos”, usamos el coeficiente de silueta, en la que una mayor puntuación del coeficiente de silueta corresponde a un modelo con *clusters* mejor definidos.

Figura 13: Coeficientes de silueta aplicados a la separación que obtiene K-medias con diferentes números de *clusters*.



En la Figura 13 vemos que la separación con 2 *clusters* es la que “mejor separa” a los datos, seguida de la separación con 3 *clusters*. Para un mayor número de *clusters*, la puntuación es considerablemente menor, lo que implica que no es razonable tomar cuatro o más *clusters*. Además, como nos interesa perfilar a los usuarios no solo por su nivel de odio, que es lo que pasaría si escogiéramos 2 *clústeres*, finalmente optamos por separar los usuarios en 3 *clusters*. Este número de agrupamientos ya habíamos podido preverlo anteriormente, con la información de la densidad bivariada de las [Figura 7](#) y [9](#).

Esta clasificación nos da, como ya hemos mencionado, un grupo en el que los usuarios se caracterizan por tener poca relación con el discurso de odio, aquellos que se caracterizan por emitir más odio y aquellos que publican más contenido upstander. A partir de ahora, llamaremos a estos grupos como los **usuarios neutros, haters y upstanders**, respectivamente. El primer grupo contiene 734 usuarios; el segundo, 1390; y el tercero, 598. Que haya más odiadores es razonable, debido a que hemos tomado una muestra de tweets a partir de palabras clave relacionadas con el discurso de odio.

6 Clasificador de usuarios en función del perfil

6.1. Modelado por características del perfil

El objetivo de esta sección es, dada la clasificación obtenida en la sección anterior, poder predecir si un autor es neutro, hater o upstander, conociendo únicamente la información de su perfil, como su nombre de usuario, el número de seguidores que tiene, a cuánta gente sigue, cuánto tiempo lleva en Twitter, etc. El propósito es doble: no solo se busca desarrollar una herramienta útil para futuros análisis (véase el [Capítulo 8](#) y el [Anexo A.1.](#)), sino también determinar si tan solo el perfil de un usuario contiene la suficiente información para desvelar si éste es odiador.

Para ello, se ha extraído una colección de características del perfil y se entrenará un clasificador que, dada esta colección como input, prediga con suficiente nivel de acierto a qué grupo pertenece cada autor. Para conseguir estas características de un autor no es necesario usar la API de Twitter, sino que se pueden extraer directamente desde la interfaz de Twitter en la aplicación del móvil (si el usuario es público).

La lista completa de las características del perfil se muestra en la [Figura 5.](#)

6.1.1. Detalles sobre algunas características del perfil

En esta sección nos ocupamos de hacer hincapié en cómo hemos obtenido algunas características del perfil de los 2722 usuarios a los que estamos analizando y/o cómo varían estas características entre los distintos tipos de usuario.

Anonimidad

Analizar esta característica viene motivada por la pregunta: ¿los usuarios hater se esconden en el anonimato más que los otros tipos? Esta situación podría darse pensando que uno tiene más libertad de expresión si no sufre repercusiones porque se conozca quién es³.

3. Es pregunta se ajusta a la RGPD ya que abarca la posibilidad de analizar datos que los usuarios publican abiertamente en redes sociales.

Tabla 5: Características del perfil usadas como input del clasificador de usuarios.

Nombre	Descripción	Valores
--------	-------------	---------

CARACTERÍSTICAS GENERALES

counts	N.º tweets	Numérico
followers	N.º seguidores	Numérico
following	N.º personas que lo siguen	Numérico
ratio_followers_vs_following	Ratio de n.º seguidores / n.º seguidos	
anon_label	Anonimidad	Numérico Anónimo, parcialmente anónimo, identificable, altamente identificable
time_since_creation	Tiempo en Twitter	Numérico (en días)
listed_count		
tweet_freq	Frecuencia de los tweets	Numérico (tweets/día)
verified	Verificado	True/False
verified_type	Tipo de verificado	Azul, empresa, gobierno, ninguno
located	Ubicado	True/False
sexo	Género del usuario	Mujer, hombre, ninguno

Nombre	Descripción	Valores
--------	-------------	---------

CARACTERÍSTICAS DE LA DESCRIPCIÓN

num_cashtags	N.º cashtags	Numérico
num_hashtags	N.º hashtags	Numérico
num_mentions	N.º menciones	Numérico
num_urls	N.º urls	Numérico
description_sentiment	Sentimiento	Positivo, negativo, neutro
description_hate	Tipo de odio	Odio-ofensivo, odio-extremo, upstander, neutro

Tabla 5: Características del perfil usadas como input del clasificador de usuarios.

Nombre	Descripción	Valores
topic_descr	Tema	artes_y_cultura, negocios_y_empresarios, celebridades_y_cultura_pop, diarios_y_vida_cotidiana, familia, moda_y_estilo, cine_tv_y_video, fitness_y_salud, comida, juegos, aprendizaje_y_educación, noticias_y_preocupación_social, otras_aficiones, relaciones, música, ciencia_y_tecnología, deportes, viajes_y_aventuras, juventud_y_vida_estudiantil
personality	Personalidad	sociabilidad, apertura, conciencia, amabilidad, neuroticismo

Para etiquetar a los usuarios en función de su nivel de anonimidad, hemos usado las mismas definiciones que en el artículo *On the internet, nobody knows you're a dog: a twitter case study of anonymity in social networks* [7]. Las definiciones son las siguientes:

- **Identificable:** Una cuenta de Twitter que contiene tanto un nombre y un apellido.
- **Altamente identificable:** Una cuenta de Twitter que es identificable y contiene una referencia URL a otra cuenta de una red social que utiliza un nombre real (como Facebook o LinkedIn). Es un subconjunto del grupo Identificable.
- **Parcialmente anónima:** cuenta de Twitter que tiene un nombre o un apellido, pero no ambos.
- **Anónimo:** Una cuenta de Twitter que no contiene ni nombre ni apellido.

Para nuestro estudio, hemos decidido además que los usuarios que están verificados por Twitter entren directamente en la categoría de Altamente Identificable, pues para recibir esta calificación deben pasar por un proceso de autenticación mucho más riguroso que tener nombre, apellido y URL con nombre real [8].

Para distinguir un nombre o apellido dentro de un nombre de usuario hemos usado el banco de nombres y apellidos del INE [9] y para separar las palabras dentro de cada nombre de usuario hemos usado la librería *ekphrasis.classes.segmenter* [10] de Python. Los resultados son los siguientes.

Figura 14: Porcentaje de usuarios neutros clasificados por nivel de anonimidad.

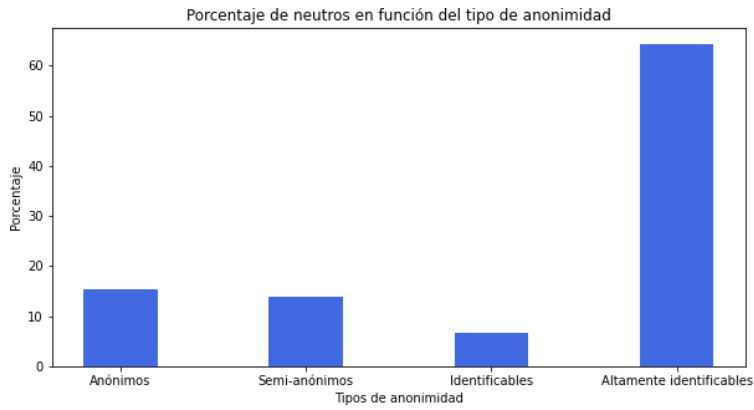


Figura 15: Porcentaje de usuarios hater clasificados por nivel de anonimidad.

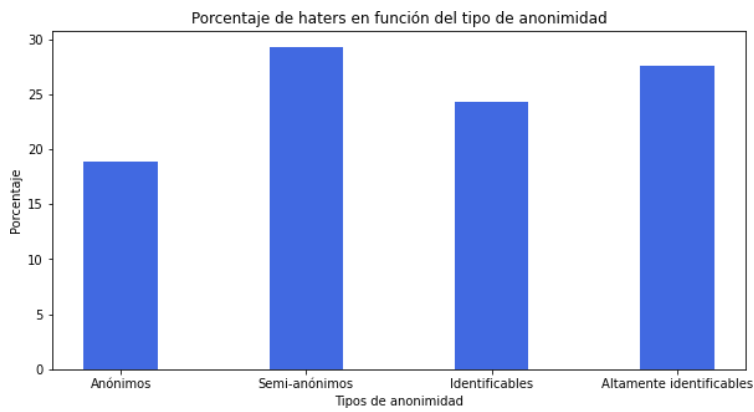
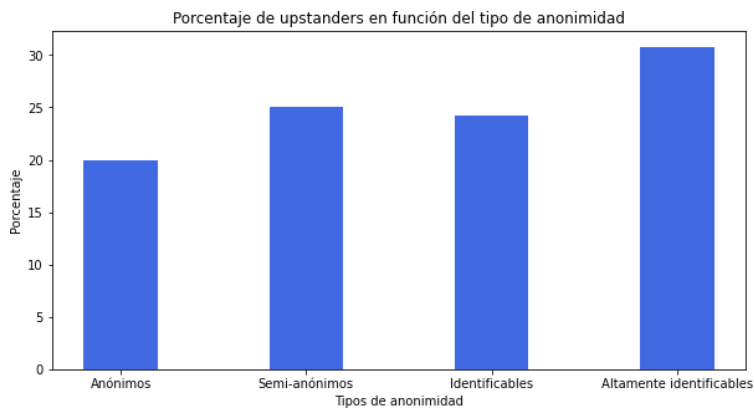


Figura 16: Porcentaje de usuarios upstander clasificados por nivel de anonimidad.

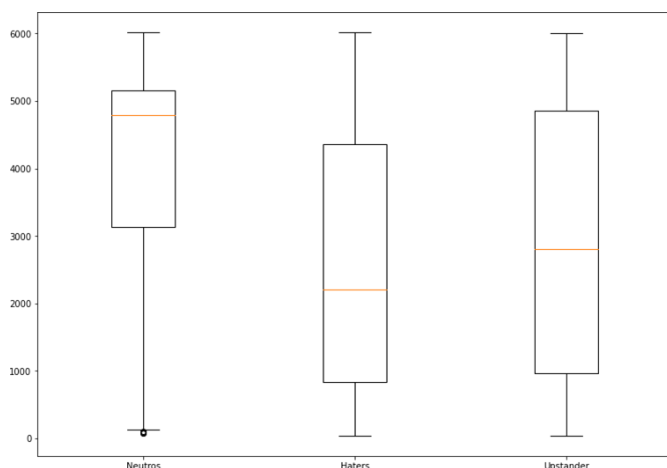


Parece obvio que las tres distribuciones son distintas puesto que los usuarios neutros se diferencian por tener muchos más usuarios altamente identificables que de otro tipo, mientras que los haters y los upstander presentan distribuciones similares. Sí que distinguimos que de entre los haters, la etiqueta de semi-anónimos es la más frecuente, mientras que entre los upstanders es la etiqueta de altamente identificados. Concluimos por tanto que los usuarios hater son los que menos se identifican en Twitter, aunque los upstander se parecen a ellos en distribución, diferenciándose en gran medida de los neutros.

Tiempo en Twitter

Otra característica de interés es el tiempo desde la creación de una cuenta, pues a priori podemos pensar que los usuarios que generan contenido odioso son más propensos a ser reportados y cerrar sus cuentas. Esta característica se representa con el número de días desde que se abrió la cuenta hasta el día en el que fueron extraídas las características de los autores.

Figura 17: Diagrama de cajas que muestra las diferencias entre el tiempo en Twitter para los distintos tipos de usuario. El eje Y corresponde al número absoluto de días. Los datos atípicos han sido descartados para una mayor claridad.

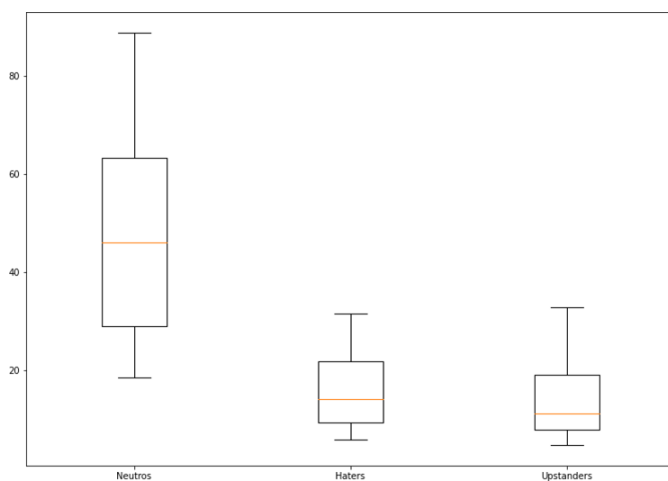


La Figura 17 nos muestra que los usuarios neutros se diferencian mucho más de los otros dos tipos pues en general llevan más tiempo en Twitter. Los tiempos desde la creación de las cuentas haters y upstanders varían más ya que el primer y tercer cuartil (la base superior e inferior de cada caja) están más alejados y lo que es más importante, la mediana de los haters es significativamente más baja que la de los upstanders.

Frecuencia de los tweets

La frecuencia de los tweets, estimada a partir de lo que conocemos del perfil, es calculada como número total de tweets dividido entre los días desde la fecha de creación. Esto nos da la media de tweets por día desde que el usuario creó la cuenta. En realidad esta es una medida que no tiene en cuenta mucha información, como que durante algunos periodos el usuario puede estar más activo que en otros, pero estos datos no se pueden extraer a partir únicamente del perfil, por lo que nos contentamos con conocer el ratio entre número de tweets por tiempo en Twitter.

Figura 18: Diagrama de cajas que muestra las diferencias entre la frecuencia en Twitter para los distintos tipos de usuario. El eje Y corresponde al ratio calculado como número total de tweets dividido entre los días desde la fecha de creación. Los datos atípicos han sido descartados para una mayor claridad.



La Figura 18 muestra la mayor distinción hasta el momento de usuarios neutros versus haters y upstanders. Los neutros son los que más tweets publican de media por día, llegando a alcanzar hasta 80 tweets por día. En contraposición, los haters y upstander, cuyos diagramas son bastante similares, no llegan a alcanzar 40 tweets por día, pero en contraposición con la variable tiempo la mediana de las haters es sensiblemente mayor que la de los upstander.

Ubicación

La etiqueta de ubicado es binaria: un “sí” significa que la ubicación del perfil permite localizar al usuario y un “no” que no se puede conocer nada sobre su ubicación. La motivación para analizar esta característica va de la mano con la cualidad de anonimato: podemos imaginar que los usuarios que no deseen ser reconocidos intentarán que tampoco se conozca dónde viven o de dónde son⁴.

Estas etiquetas han sido obtenidas manualmente puesto que, en base a nuestro conocimiento, no existe ninguna herramienta que nos permita realizar esta clasificación. Hemos considerado marcar la etiqueta de ubicado a “sí” dada la ubicación, que permita inferir mínimamente dónde se encuentra un usuario. Por ejemplo, regiones muy amplias (Latinoamérica, Europa), países, nombres ambiguos (“El Llano”, que puede indicar un distrito de Gijón o una localidad en Panamá), las ubicaciones propias de Twitter (“ÜT”, seguido de coordenadas de latitud y longitud), códigos postales, avenidas y ríos. Las localizaciones que no entren en esta clasificación o que estén en blanco son marcadas con “no”.

Adicionalmente, los usuarios verificados con la etiqueta de empresa o de gobierno, las cuales no corresponden a usuarios individuales que estén ubicados en un lugar concreto, fueron marcadas con “sí”.

Los resultados son los siguientes: Haters: 63% ubicados, Upstanders: 72% ubicados, Neutros: 84% ubicados. Los haters son los usuarios menos localizados, seguidos de los upstander.

La personalidad

Aunque existen varios modelos para describir la personalidad, una de las teorías más investigadas y aceptadas es el modelo de los cinco grandes rasgos (o Big Five), formado por la sociabilidad, la apertura a nuevas experiencias, la conciencia, la amabilidad y el neuroticismo [11]. Se ha demostrado que cada una de estas personalidades deja una huella en las redes sociales [12], por lo que es interesante asociar a cada usuario un rasgo de este modelo. Para lograrlo solo con la información del perfil, clasificamos la descripción con el modelo `Minej/bert-base-personality` de la plataforma HuggingFace, habiéndolo traducido anteriormente con la herramienta `Helsinki-NLP/opus-mt-es-en`, también de HuggingFace.

El tema de la descripción

El modelo de CardiffNLP que clasifica un texto por su tema ha sido entrenado expresamente con contenido de Twitter y, además, los temas se seleccionaron basándose en las tendencias de Twitter, con el objetivo de que fueran amplios y generales y consistieran en clases como: arte y cultura, música o deportes.

⁴ Esta pregunta se ajusta a la RGPD ya que abarca la posibilidad de analizar datos que los usuarios publican abiertamente en redes sociales, en este caso con fines científicos para dar seguimiento al proceso de descubrimiento de los patrones generales de comportamiento de usuario en Twitter, no porque de alguna forma se almacenen en este proyecto las ubicaciones concretas de usuarios.

Por ello, es totalmente indicado usarlo en este trabajo para extraer el tema principal de la descripción del perfil.

El género

Dados los nombres de usuario (el personalizado o el identificativo⁵), se busca si contienen un nombre de mujer o de hombre, según el banco de nombres del INE [\[9\]](#). Si no contienen ninguno o contienen a ambos, no se asigna ningún género al usuario.

6.2. Clasificación con árbol de decisión

Tras añadir estas características a la matriz de datos, se entrena el clasificador basado en un árbol de decisión, usando la librería de Python `xgboost`. Se implementa “oversampling” en el entrenamiento, para balancear las clases, y se elige un tamaño de prueba del 20%.

Los métodos basados en árboles son útiles para la interpretación, por ello los elegimos por encima de otros modelos para este proyecto en concreto.

6.2.1. Funcionamiento de un Árbol de Decisión

En breve, este método funciona dividiendo recursivamente el espacio de predicción en regiones más pequeñas y asignando una clase a cada región. Más detalladamente, comienza con el conjunto de datos completo en el nodo raíz y evalúa distintos puntos de división para cada característica y selecciona aquel que minimiza o maximiza una función objetivo.

En este caso, hemos elegido `reg:squarederror` como función objetivo, puesto que con ella se obtenía la mejor puntuación. Se caracteriza por minimizar la suma de los errores las diferencias al cuadrado entre los valores objetivo reales y los valores predichos. Tras la primera división, el conjunto de datos se divide en dos subconjuntos, y se repite el proceso recursivamente hasta que se cumplan determinados criterios de parada. Al final, cada nodo terminal del árbol contiene un valor constante que permite decidir a qué clase pertenece esa partición, y es la media de los valores objetivo de las muestras dentro de ese nodo hoja. Por ejemplo, como la etiqueta “hater” se ha codificado como 0 y la etiqueta “neutro”, como 1, si el valor del nodo hoja es 0.7, entonces la clase resultante es “neutro”, por estar más cerca del 1 que del 0.

5. En Twitter, hay dos nombres de usuario: el identificativo, que es único y está compuesto por caracteres estándar (por ejemplo, “@irene104”) y el personalizado, que pueden compartir varias personas y que permite insertar emoticonos u otros caracteres especiales (por ejemplo, “★ Irene ★”).

6.2.2. Resultado

La puntuación del clasificador se muestra en la Tabla 6.

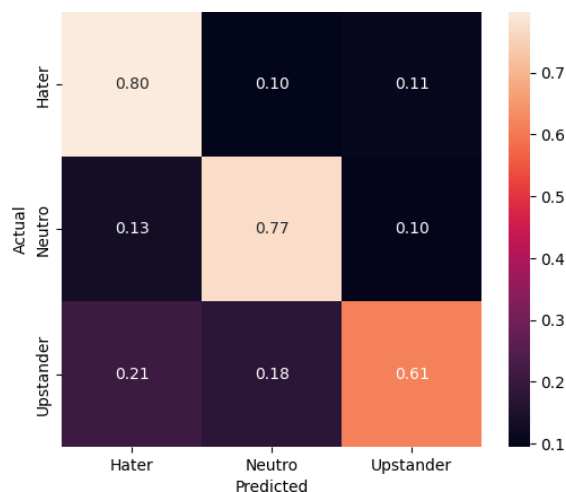
Tabla 6: Puntuación del clasificador de usuarios en función de su perfil.

Medida	Puntuación
Accuracy	0,7560
F1 (weighted)	0,7571
F1 (macro)	0,7384

Estos valores demuestran que efectivamente **existe una relación significativa entre las características del perfil y el tipo de usuario**, dado que se alcanza un 0.75 de precisión en la clasificación, en comparación con el 0.33 que se obtendría si clasificáramos al azar. Esta observación es importante: nos indica que no hace falta recurrir al registro de tweets de un usuario para saber de qué tipo son, con los datos de la cuenta ya tenemos una brújula fiable a la que consultar. Cabe mencionar que, aunque se ajustaron los hiperparámetros del estimador con la técnica de 10-fold Cross-validation, las puntuaciones no mejoraron.

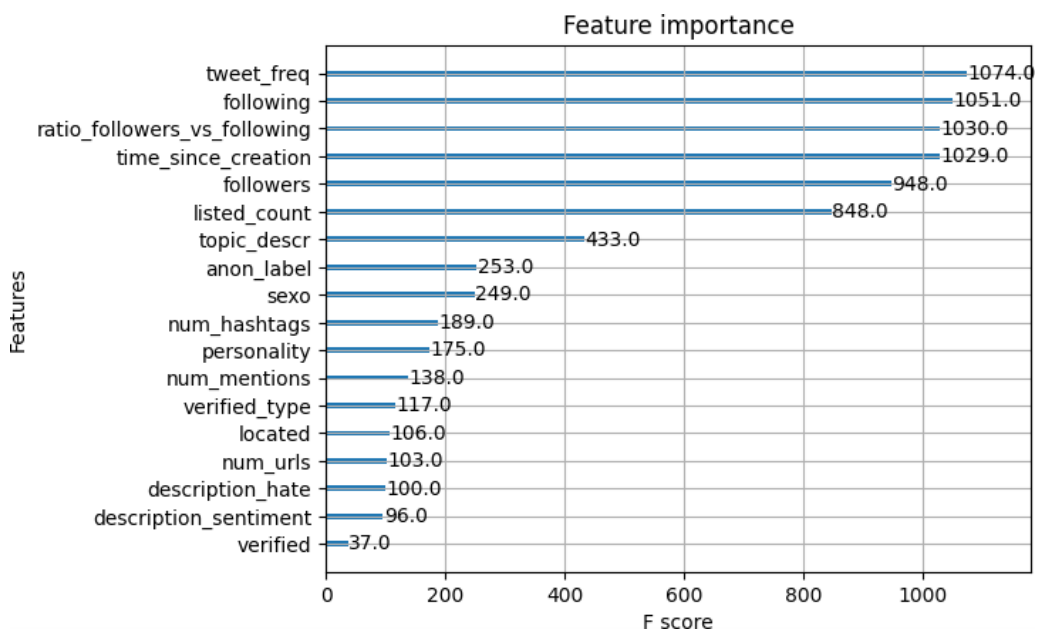
La Figura 19 muestra la matriz de confusión. Se observa que los odiadores y los neutros son los que mejor se clasifican. Aunque los upstander se queden rezagados, se los reconoce un 66% de las veces: un gran aumento respecto al 49% de la primera versión del clasificador. El hecho de que no lleguen a clasificarse igual de bien puede deberse a su parecido con los haters en términos de las características del perfil.

Figura 19: Matriz de confusión obtenida para las clases hater, neutro y upstander.



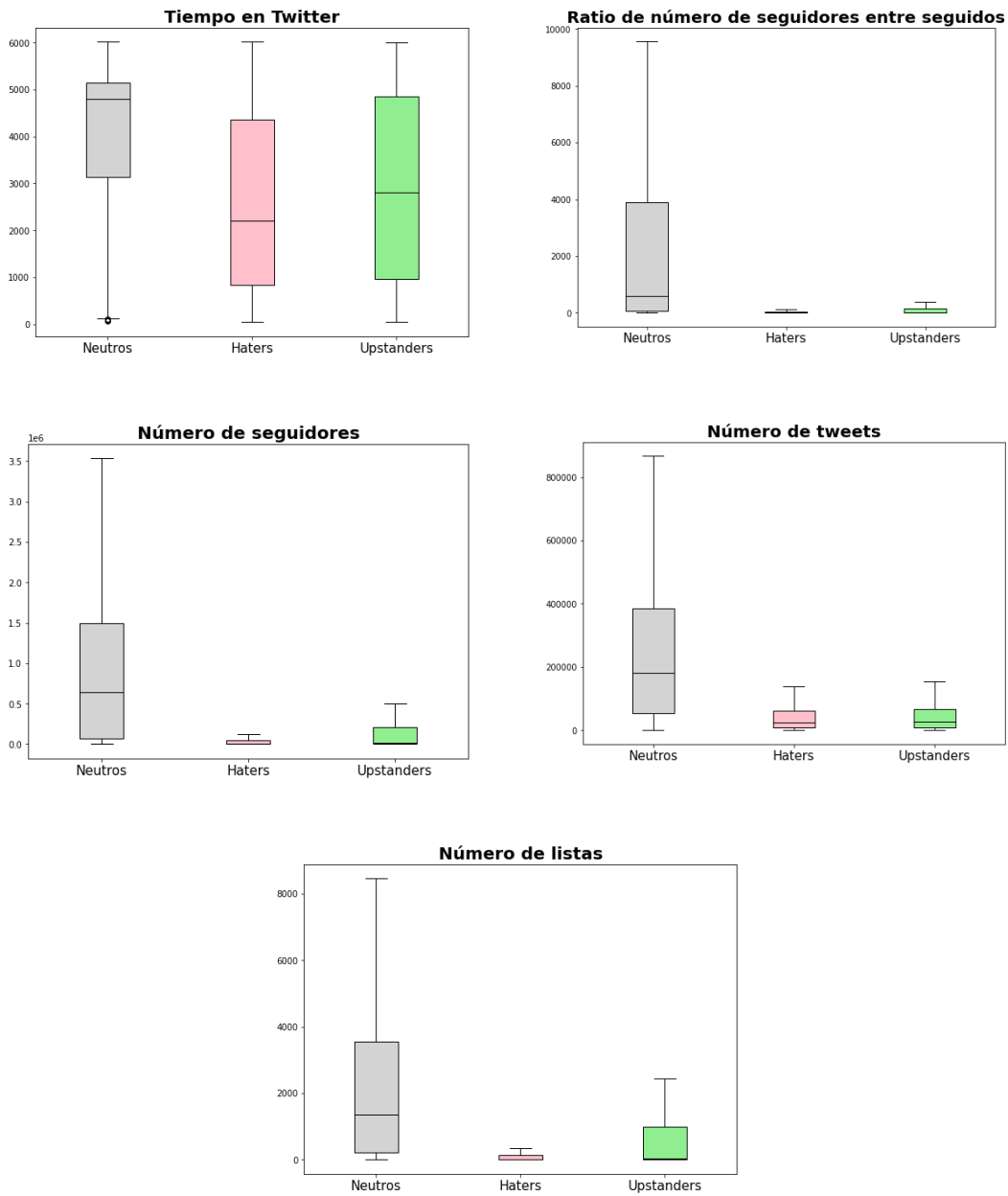
Como el clasificador que hemos implementado es un árbol de decisión, podemos extraer fácilmente la importancia que éste le ha dado a cada una de las características del perfil.

Figura 20: Importancia de las características del perfil identificadas en el modelo.



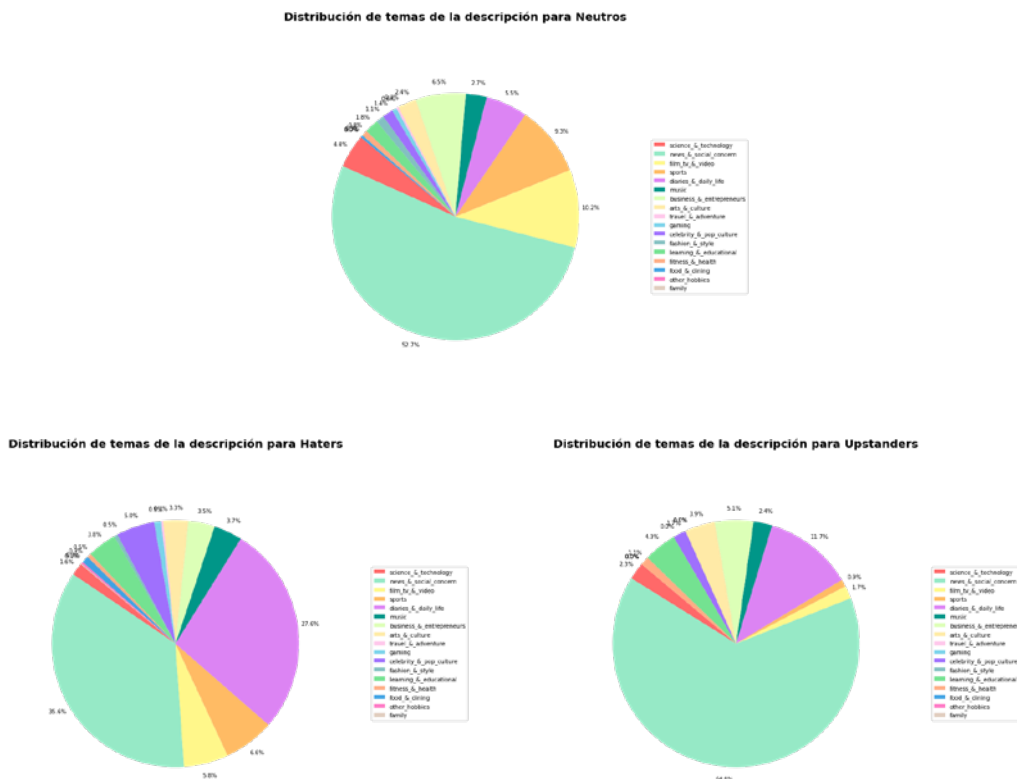
La Figura 20 nos muestra el nivel de importancia de cada característica. Está medido con F score, un medidor que depende del número de veces que aparezca en el árbol para segmentar dos caminos y de la altura a la que se usen. Cuanto más alto sea, más importancia denota. Por tanto, las características más relevantes a la hora de clasificar son, por orden, la frecuencia con la que se escriben tweets, el número de seguidores, el ratio entre seguidores y seguidos, el tiempo en Twitter, el número de listas a las que un usuario está unido y el tema de la descripción. Cómo se distribuye cada una por tipo de autor se muestra en las figuras siguientes.

Figura 21: Diagrama de cajas de las cinco características más importantes del modelo por tipo de autor. Se ha elegido no graficar los valores atípicos por cuestión de claridad.



Vemos que los atributos más notables son los que distinguen si un usuario tiene presencia en la red, está activo y si lleva mucho tiempo. Y por el análisis realizado anteriormente, es posible que estos factores escindan entre neutros y no neutros. Los neutros son los que tienen construidas plataformas grandes y los hater/upstander no.

Figura 22: Gráfico circular de la sexta característica más importante por tipo de usuario.



Las características anteriores no distinguían claramente entre los haters y los upstanders, sin embargo, el tema de la descripción sí lo hace. Se observa que, aunque el tema predominante en ambos grupos es el de noticias y preocupación social, y el segundo más predominante es diarios y vida cotidiana, los porcentajes son notablemente diferentes. Mientras que un 35,6% y un 27,6% de los usuarios odiadores hablan del primer y segundo tema, respectivamente, son un 64,8% y un 11,7% de upstanders quienes mencionan el primer y segundo tema, respectivamente. Es decir, **hay muchos más upstanders que tratan noticias y preocupación social en su descripción que haters**, y menos upstanders que mencionan diarios y vida cotidiana.

En resumen, las cinco características más relevantes distinguen entre los usuarios neutros y los no neutros, mientras que la sexta característica diferencia entre los haters y los upstanders. Las características restantes proporcionan detalles adicionales para cada una de las ramas del árbol de decisión.

7 Estudio de los usuarios a partir de sus tweets

El objetivo de esta sección es usar la clasificación de los usuarios y el registro de sus últimos 1000 tweets para estudiar el perfil de las personas involucradas en el discurso de odio en Twitter. En concreto, se han recopilado un conjunto de preguntas específicas que pretenden analizar los patrones ocultos en los tweets de los usuarios. Esta sección explica cómo se han respondido a estas preguntas y las conclusiones que se extraen.

7.1. ¿Un usuario habla de temas diferentes cuando emite odio en comparación con cuando no lo hace?

Una primera cuestión a analizar es si los autores, cuando escriben mensajes de odio, los temas de los que hablan difieren de cuando no lo hace. Por ejemplo, un usuario podría hablar usualmente de lo que hace durante el día, sin tener esto nada que ver con el discurso de odio, pero luego, al debatir temas políticos o deportivos, puede alterarse y escribir mensajes ofensivos. Esta pregunta se puede generalizar a si el tema del que trata un tweet (vida cotidiana, noticias, deportes, etc.) depende de qué tipo de relación tenga con el discurso de odio (odio, neutro, upstander). Más formalmente, en un lenguaje estadístico, queremos saber si la distribución que sigue la variable aleatoria “tema” es diferente bajo tres condiciones diferentes: que el tweet sea de odio, neutro o upstander. Para ello, existe un método estándar de enfrentarse a este problema, denominado contraste de homogeneidad.

El primer paso de este método consiste en construir la **tabla de frecuencias** de cada usuario, la cual alberga el número de tweets por tema⁶ y por tipo (con tipo nos referimos aquí a si es neutro, de odio o upstander). La [Tabla 7](#) es un ejemplo de un usuario inventado que solo habla de noticias y música y que emite tantos tweets neutros como upstander como hater con ambos temas, sin hacer distinción alguna. Por tanto, podríamos deducir que la distribución que sigue la variable aleatoria “tema” de este usuario es la misma bajo las tres condiciones diferentes. Pero si, por ejemplo, un usuario generara la [Tabla 8](#), sabríamos entonces que cuando emite odio está hablando sobre noticias y cuando no, es que habla sobre música. Y por ende, podríamos asegurar que la variable “tema” de este autor se distribuye de manera diferente dependiendo de la variable “tipo”. Sin embargo, ¿qué se podría decir de la [Tabla 9](#)? no sigue ningún patrón evidente. Parece que habla de múltiples temas, a veces solo con tono neutro, pero otras veces con todos los tonos, por lo que a simple vista no se puede sacar ninguna conclusión. Además, en esta tabla el caso más habitual, según lo observado al realizar este análisis.

⁶ Los temas han sido clasificados con el modelo de CardiffNLP [\[5\]](#), entrenado específicamente con un dataset de tweets.

Es por este último ejemplo que se necesita el siguiente paso: resumir toda la información de la tabla de frecuencias en un único número que permita dar una respuesta clara a la hipótesis de que no haya diferencias significativas entre las tres distribuciones. Este número, en la teoría de los contrastes estadísticos, se denomina **p-valor**, y se obtiene a través de una fórmula que toma como entrada las frecuencias y la dimensión de la tabla de frecuencias. Cuanto más pequeño sea el p-valor, más improbable es que la distribución de los temas no dependa del tipo de tweet. De hecho, el límite estándar a partir del cual se rechaza esta hipótesis es 0,05. En otras palabras, si el p-valor es menor que 0,05, podemos afirmar con casi total seguridad de que los temas de los que habla un autor sí que dependen del tipo de odio que esté emitiendo.

Los resultados son los siguientes. Resulta que tras aplicar el contraste a cada usuario, la hipótesis se rechaza un 0.56% de las veces. **En conclusión, más de la mitad de los usuarios, al relacionar su tweet con el discurso de odio, cambian de tema.**

Tabla 7: Tabla de frecuencias de un usuario que solo habla de noticias y música, y cuando lo hace, el tweet puede ser de cualquier tipo.

Tópico	Tipo		
	Neutro	Odio	Upstander
artes y cultura	0	0	0
negocios y empresarios	0	0	0
celebridades y cultura pop	0	0	0
diarios y vida cotidiana	0	0	0
familia	0	0	0
moda y estilo	0	0	0
cine tv y vídeo	0	0	0
fitness y salud	0	0	0
comida	0	0	0
juegos	0	0	0
aprendizaje y educación	0	0	0
noticias y preocupación social	100	100	100
otras aficiones	0	0	0
relaciones	0	0	0
música	30	30	30
ciencia y tecnología	0	0	0
deportes	0	0	0
viajes y aventuras	0	0	0
juventud y vida estudiantil	0	0	0

Tabla 8: Tabla de frecuencias de un usuario que cuando emite odio está hablando sobre noticias y cuando no, es que habla sobre música.

Tópico	Tipo		
	Neutro	Odio	Upstander
artes y cultura	0	0	0
negocios y empresarios	0	0	0
celebridades y cultura pop	0	0	0
diarios y vida cotidiana	0	0	0
familia	0	0	0
moda y estilo	0	0	0
cine tv y vídeo	0	0	0
fitness y salud	0	0	0
comida	0	0	0
juegos	0	0	0
aprendizaje y educación	0	0	0
noticias y preocupación social	0	100	0
otras aficiones	0	0	0
relaciones	0	0	0
música	30	0	0
ciencia y tecnología	0	0	0
deportes	0	0	0
viajes y aventuras	0	0	0
juventud y vida estudiantil	0	0	0

Tabla 9: Ejemplo de un usuario que habla de múltiples temas, sin seguir ningún patrón aparente de cuáles de estos le sugieren escribir un tipo concreto de tweet.

Tópico	Tipo		
	Neutro	Odio	Upstander
artes y cultura	3	0	0
negocios y empresarios	5	5	4
celebridades y cultura pop	55	0	0
diarios y vida cotidiana	199	59	24
familia	20	2	15
moda y estilo	0	0	0
cine tv y vídeo	3	0	4
fitness y salud	7	1	0
comida	2	0	0
juegos	5	100	7
aprendizaje y educación	0	0	0
noticias y preocupación social	55	68	1
otras aficiones	1	0	0
relaciones	21	10	5
música	38	39	43
ciencia y tecnología	2	0	0
deportes	0	0	0
viajes y aventuras	7	3	0
juventud y vida estudiantil	0	0	0

7.1.1. ¿Un hater discrimina a un único colectivo o a varios?

Sería interesante conocer si los usuarios odiadores suelen apuntar a un único objetivo o si, por el contrario, diversifican su odio. Por ejemplo, ¿un usuario únicamente emite comentarios degradantes de corte racial o también discrimina por religión y sexo? Pero además, si miramos al conjunto de todos los autores, ¿cuál es la tónica general? ¿La mayoría tiende a odiar solo a un colectivo o a varios?

En primer lugar, para llevar a cabo este análisis se ha de clasificar cada tweet en función de a qué entidad ataca. A priori nos gustaría usar la clasificación del proyecto original [2]: odio hacia personas gitanas, judías, migrantes, musulmanas y hacia otras minorías culturales y etnias religiosas. Pero, desgraciadamente,

no existe ninguna herramienta que maneje estas clases específicas y desarrollar una *ad hoc* sería tan costoso que no entraría dentro del alcance de este trabajo. Sin embargo, sí que existe el modelo `cardiffnlp/twitter-roberta-base-hate-multiclass-latest`, integrado dentro del amplio repertorio de CardiffNLP [5], que clasifica tweets odiadores en una de las siguientes clases:

Tabla 10: El modelo de CardiffNLP clasifica un tweet odiador en una de estas clases.

Sexismo	Discapacidad	Racismo
Religión	Orientación sexual	Otros

Además, está diseñado para manejar no textos cualesquiera, sino tweets, con lo que se adapta perfectamente a nuestro contexto. La desventaja es que solo procesa información en inglés, por lo que antes se deben traducir todos los tweets⁷.

Una vez ya están traducidos y clasificados todos los tweets de usuarios hater, los agrupamos por autor. Resulta que del total de haters, obtenemos con el clasificador de Cardiff que un 2% no escribió tweets de odio. Esto a priori puede llamar la atención, pero es muy posible que se deba a que el clasificador no busca el mismo tipo de odio que HaterBert y entonces ocurra que algunos tweets de odio pasen desapercibidos, además de que al traducir se puede perder contexto. Se estima que una cantidad del 2% no es significativa en este contexto.

Del 98% restante de usuarios se obtienen los siguientes resultados: El 44% de los autores emiten exclusivamente un tipo de discriminación. La categoría más común para este tipo de tweets es “otros”. Esto puede explicarse nuevamente por el hecho de que estos autores han sido seleccionados por emitir odio contra grupos como personas gitanas, judías, migrantes, etc., categorías que no se ajustan a ninguna específica de la Tabla 10. Un 37% emiten dos tipos de discriminación. Un 14% emiten tres tipos de discriminación. Un 4%, 0,7% y 0,1% emiten cuatro, cinco y hasta 6 tipos de discriminación, respectivamente. Se puede ver en más detalle en las Figuras 23, 24 y 25.

En conclusión, **la mayoría de usuarios se centran en discriminar a un único colectivo**. Aunque podría ser que la categoría “otros” englobara a varios subgrupos, unos no contemplados en la clasificación de la Tabla 10.

7. Con el modelo Helsinki-NLP/opus-mt-es-en de HuggingFace.

Figura 23: Gráfica que muestra en cada barra el porcentaje de cada tipo de discriminación que emite cada usuario, para usuarios que emiten hasta 4 tipos (los nombres de los usuarios están difuminados por motivos de protección de datos).

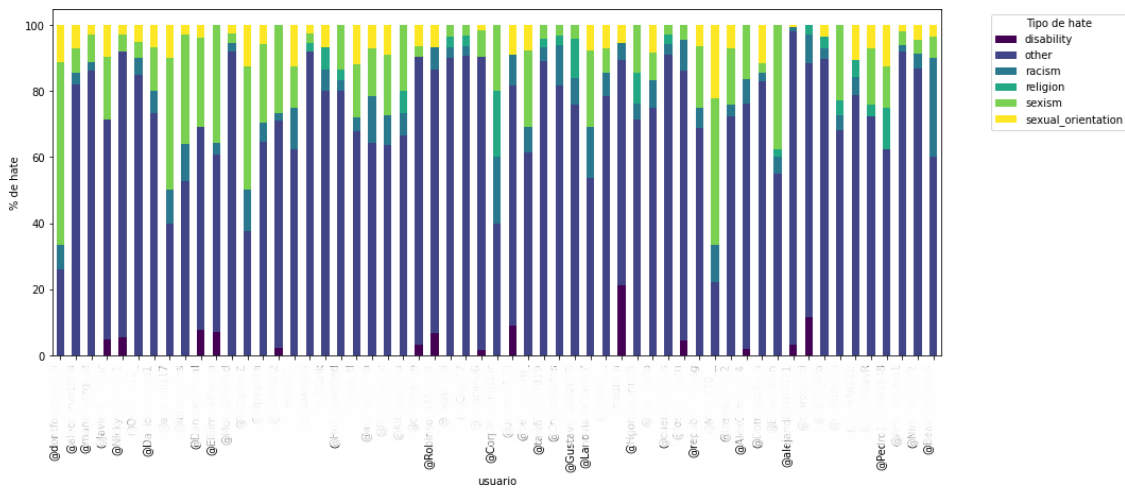


Figura 24: Gráfica que muestra en cada barra el porcentaje de cada tipo de discriminación que emite cada usuario, para usuarios que emiten hasta 5 tipos (los nombres de los usuarios están difuminados por motivos de protección de datos).

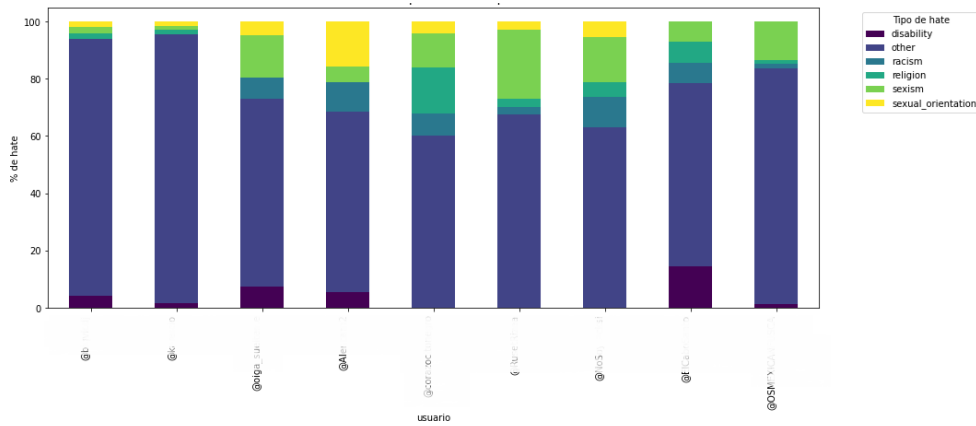
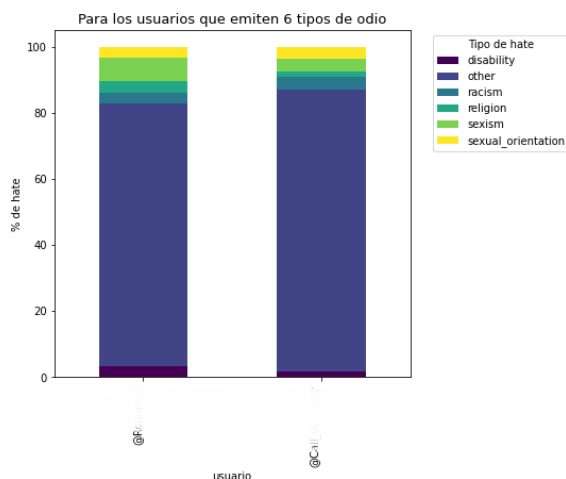


Figura 25: Gráfica que muestra en cada barra el porcentaje de cada tipo de discriminación que emite cada usuario, para usuarios que emiten hasta 6 tipos (los nombres de los usuarios están difuminados por motivos de protección de datos).



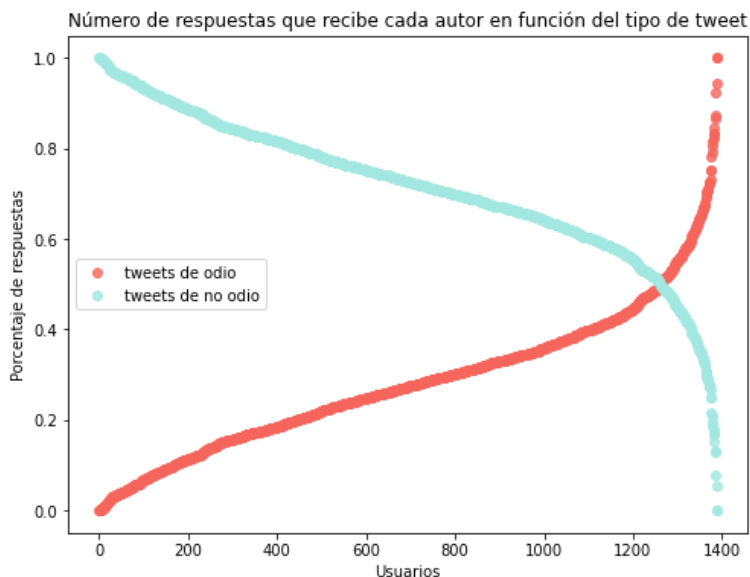
7.1.2. Sobre los odiadores: ¿son sus tweets de odio más incendiarios que los que no son de odio?

Se va a usar un valor bastante natural que mide cómo de “incendiario” es un tweet: el número de respuestas.

Primero, identificamos a los usuarios clasificados como haters y calculamos las medias correspondientes de respuestas para los tweets de odio (ofensivo y extremo) y no odio (neutro y upstander), agrupándolos por autor. Luego, como no nos interesa el valor absoluto de los promedios sino cómo se comparan entre ellos, normalizamos estos valores a su porcentaje relativo.

A continuación, graficamos el promedio de respuestas porcentual de los tweets de no odio versus el promedio de los de odio ([Figura 26](#)), habiendo ordenado a los usuarios por este último.

Figura 26: Por ejemplo, si el 400-ésimo usuario recibe de media X número de respuestas en sus tweets de odio e Y respuestas en sus tweets de no odio, entonces por la gráfica sabemos que X representa el 20 % de $X + Y$ e Y representa el 80 % de $X + Y$.



Se observa que la línea roja no sobrepasa el 50% hasta aproximadamente el usuario 1250-ésimo. Es decir, de los 1390 haters, solo unos 140 suelen recibir más respuestas en sus tweets de odio en comparación con los de no odio. Además, la gráfica nos permite apreciar que, de los 1250 restantes, la mayoría presenta el porcentaje de respuestas de tweets de odio por debajo del 0,4. Es decir, se concluye que **en general, los tweets de odio son menos incendiarios que los de no odio.**

7.1.3. ¿Los usuarios odiadores, neutros y upstander tienen diferentes personalidades?

Según la clasificación Big 5, queremos analizar si los diferentes tipos de usuarios se corresponden con diferentes personalidades. Para ello, primero se cataloga cada tweet de cada autor con la personalidad que más se asocia a dicho texto, usando el clasificador `Minej/bert-base-personality`. Segundo, a cada autor se le asigna la personalidad que se haya encontrado en el mayor número de tweets. Tercero, se separa a los autores por tipo y por cada grupo se calcula cuántos autores hay de cada personalidad. Los resultados se muestran en la [tabla 11](#).

Tabla 11. Grupos de personalidad de cada autor.

	5 grandes rasgos				
	Sociabilidad	Neuroticismo	Amabilidad	Conciencia	Apertura
Neutros	10 (1%)	723 (99%)	0	0	1 (0,1%)
Haters	11 (0,8%)	1377 (99%)	0	0	2 (0,1%)
Upstanders	4 (0,7%)	593 (99%)	0	0	1 (0,2%)

Observamos que la mayoría de tweets escritos por la mayoría de usuarios son de carácter neurótico, aunque hay algunos que presentan más carácter social o de apertura. Sin embargo, no se aprecia ninguna diferencia entre los diferentes tipos de usuario. Por tanto, la personalidad no es indicadora de ningún tipo de usuario.

8

Grafo de conversaciones

La tarea final de este proyecto es analizar las relaciones entre los diferentes tipos de usuarios. Si las personas que contestan a los mensajes ofensivos son también personas que vierten odio a la red, si son meros observadores o si se oponen al odio. Más concretamente, este objetivo se puede desglosar en dos:

1. Los usuarios de cada tipo (hater, upstander y neutro), ¿con qué tipo de usuarios se suelen relacionar? ¿El mismo o diferente?
2. ¿Cómo son estas interacciones? ¿De apoyo, de indiferencia o de oposición?

La manera de ver esto será creando un gráfico de la red que muestre a la vez ambos puntos. Seremos capaces de ver qué usuarios se relacionan y de qué manera, si apoyándose entre ellos o posicionándose en contra.

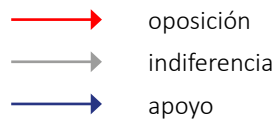
8.1. Diseño del grafo

Un grafo es una estructura formada por un conjunto de nodos que conectan entre sí a través de aristas. En este caso, cada usuario estará representado por un nodo y las aristas que los comuniquen con otros simbolizarán que uno ha contestado a un tweet del otro. Las aristas tendrán dirección, es decir, visualmente serán una flecha que parte de del autor del primer tweet y que apunta a la persona que responde a dicho tweet.

El color de los nodos representará el tipo de usuario, conforme a la siguiente leyenda.



Y el color de las aristas representará la naturaleza de la interacción entre los dos nodos que conecta.

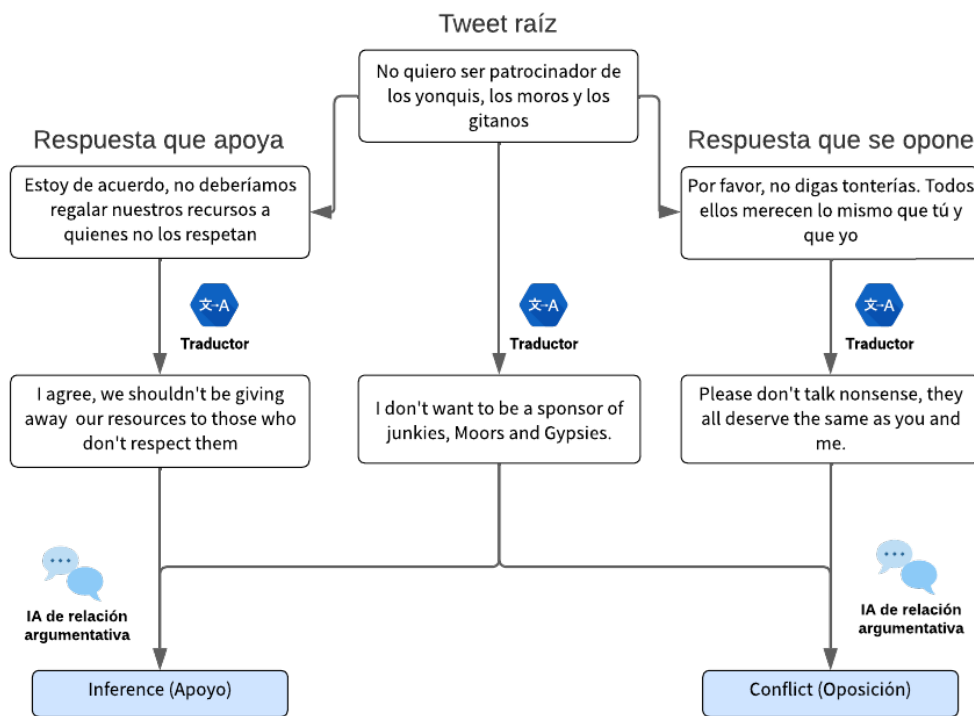


8.2. Obtención de las aristas

Primero, filtramos los tweets extraídos mediante la búsqueda por keywords (la cual se ha descrito en la [Sección 3.3.](#)), descartando los que no hayan generado una conversación. Segundo, traducimos todos los tweets, tanto los raíz como sus respuestas, para después procesarlos con una IA que dictaminará el tipo de relación argumentativa que existe entre ellos, ya que solo maneja textos en inglés.

En particular, el traductor usado es Helsinki-NLP/opus-mt-es-en, un modelo de procesamiento de lenguaje natural de la plataforma HuggingFace, al igual que el modelo de relación argumentativa, raruidol/ArgumentMining-EN-ARI-AIF-RoBERTa_L. Este último decide si la respuesta se opone (Conflict), es una reformulación (Rephrase), se infiere de lo anterior (Inference) o si no tiene relación (No-relation). Para este trabajo, entenderemos que tanto una reformulación como una inferencia son interacciones de apoyo. La [Figura 27](#) detalla un ejemplo de cómo se han usado estas herramientas.

Figura 27: Ejemplo del proceso de obtención de la relación argumentativa entre dos respuestas y el mensaje al que responden. Este último es un extracto de un tweet más largo que se encuentra en nuestra base de datos, a diferencia de las respuestas, que simplemente se han ideado aquí como ejemplo.



8.3. Obtención de los nodos

El siguiente paso es clasificar todos los usuarios involucrados en las conversaciones en hater, upstander o neutro. Como no se dispone de una muestra de sus tweets, no podemos discernir su tipo a partir del porcentaje de odio que emitan. Sin embargo, sí se llegaron a descargar los perfiles de cada usuario, por lo que se decide usar el clasificador descrito en el [Capítulo 6](#). Es decir, primero se seleccionan las características explícitas del perfil (número de seguidores, de seguidos, de tweets...), después se deducen las implícitas (anonimidad, frecuencia en Twitter, etc.) con los mismos algoritmos descritos en la [Sección 6.1.1](#), y finalmente se pasa el clasificador de usuarios.

8.4. Resultados

La [tabla 12](#) muestra el porcentaje de tipos de nodo y de arista obtenidos por categoría de odio. Se observa que, en general, la mayoría de aristas son de apoyo y la mayoría de nodos son de haters. Lo primero indica que los usuarios tienden a relacionarse con personas afines a sus principios, lo segundo se puede deber a que estas conversaciones contenían palabras relacionadas con el discurso de odio. Aunque es cierto que el porcentaje de odiadores es muy alto (en torno al 90%), esto quizá se debe a que el clasificador de usuarios etiqueta como neutros a los usuarios con presencia en la plataforma, cuando en realidad hay más neutros que no tienen tanto peso.




Tabla 12. porcentaje de tipos de nodo y de arista obtenidos por categoría de odio.

	Tipos de nodo			Tipo de arista Categoría		
	Neutros	Haters	Upstanders	Oposición	Apoyo	Indiferencia
Globales	3%	90%	7%	5%	46%	49%
Minorías culturales y etnias religiosas	3%	91%	6%	5%	54%	41%
Personas gitanas	1%	93%	6%	10%	56%	34%
Personas judías	2%	92%	6%	7%	56%	37%
Migrantes	3%	90%	7%	6%	51%	42%
Personas musulmanas	3%	91%	6%	6%	54%	40%

A continuación se incluyen algunos fragmentos de grafos obtenidos:

Figura 28: Ejemplos de relaciones entre usuarios hater y upstander, junto a las conversaciones de las que se obtuvieron estos grafos. Los nombres están anonimizados por cuestiones de protección de datos.





 @us1	un día más yendo de empalme a trabajar porque golfo se nace pero hay que costearse la vida gitana
 @us2	Eres leyenda
 @us1	me están saliendo las agujetas de to la semana de no descansar bro

Dos haters se apoyan entre sí. Esto proviene de la conversación que se muestra a la derecha (los nombres de los usuarios están difuminados por motivos de protección de datos).

Primero, @us1 emite un estereotipo degradante hacia el colectivo gitano, al cual @us2 responde, validándolo y elogiándolo, y @us1 concluye contestando a @us2 también de manera positiva. Los nombres están anonimizados por cuestiones de protección de datos.



	Lxs alfrely: nos gustaría ver más representación lgbt y diferente minorías en la televisión
 @us3	Z0wl: la cultura de los gitanos es el crimen y por eso deberíamos eliminarlos
	Tu comparación tiene cero sentido
 @us5	Porque esa no fue la comparación que quería hacer

Un hater se opone a un upstander. Esto proviene de la conversación que se muestra a la derecha (los nombres de los usuarios están difuminados por motivos de protección de datos).

Primero, @us3 manifiesta su opinión sobre un asunto controvertido (originado por @us4, un usuario popular de la red), a la cual @us5 se opone.

La [Sección A.4. del Anexo A](#) incluye un tutorial sobre cómo visualizar y explorar los grafos obtenidos. Al visualizarlos, se ve inmediatamente que los haters se relacionan sobre todo con otros haters, y por tanto podemos concluir que no es tanto que los haters interaccionen con víctimas de su discriminación, sino que se comunican con personas de mentalidad afín.

A modo de broche final, se puede considerar que cada uno de los grafos representa un nicho de odiadores que expresan odio hacia un determinado colectivo. En caso de querer analizar en el futuro a otro conjunto específico, entonces se tendría que realizar una nueva búsqueda de tweets con palabras clave relacionadas con otro tipo de discriminación.

9 Resumen general de logros

El proyecto desarrollado ha logrado avances significativos en la clasificación y análisis de autores en Twitter, con un enfoque particular en la detección de mensajes de odio y contranarrativa.

Clasificación de Autores en Twitter: El proyecto ha establecido una metodología efectiva para la clasificación de autores en Twitter en función de la cantidad de mensajes de odio y contranarrativa que emiten. Esto ha permitido analizar los atributos específicos de cada tipo de autor, como la cantidad de tweets, la antigüedad en la plataforma y los temas de conversación.

- **Predicción del Tipo de Autor:** Se ha desarrollado un modelo predictivo capaz de determinar el tipo de autor basado únicamente en la información del perfil del usuario en Twitter. Este enfoque no solo ha proporcionado una herramienta útil para el análisis requerido en el estudio, sino que también ha revelado la capacidad de la información del perfil para predecir la emisión de mensajes de odio.
- **Perfilado de Usuarios:** El análisis de patrones y atributos del perfil y los tweets ha permitido caracterizar a los autores que emiten una gran cantidad de odio. Este proceso de perfilado ha proporcionado información valiosa para comprender mejor el comportamiento de los usuarios en la red social.
- **Grafo de Conversaciones:** Se ha investigado la naturaleza de las interacciones entre autores que emiten mensajes de odio, analizando si actúan de manera individual o forman parte de colectivos de odiadores. Este análisis ha arrojado luz sobre la dinámica de las conversaciones en Twitter y la relación entre los emisores de odio y otros usuarios.
- **Desafíos y Superación:** El proyecto ha enfrentado desafíos significativos, como el cese del acceso gratuito a la API de Twitter, lo que ha requerido adaptaciones en los objetivos y metodologías. A pesar de estas dificultades, se han implementado soluciones efectivas para llevar a cabo la extracción de datos y el análisis de los mismos.

9.1. Conclusiones

En resumen, el proyecto se ha centrado en el desarrollo de un clasificador de usuarios de Twitter basado únicamente en las características de sus perfiles, con el objetivo de predecir si un autor es neutro, hater o upstander. Se ha demostrado que las características del perfil contienen información relevante para realizar esta clasificación con un nivel de precisión considerable.

En primer lugar, se ha realizado un análisis detallado de diversas características del perfil de los usuarios, como la anonimidad, el tiempo en Twitter, la frecuencia de tweets, la ubicación, la personalidad y el tema de la descripción. Se observó que las distribuciones de estas características difieren entre los usuarios neutros y los no neutros, y entre los haters y los upstanders.

Posteriormente, se implementó un clasificador basado en árboles de decisión, que logró una precisión del 75%, superando significativamente la precisión esperada al azar (33%). Se identificó que las características más importantes para la clasificación fueron la frecuencia de tweets, el número de seguidores, el ratio entre seguidores y seguidos, el tiempo en Twitter, el número de listas a las que un usuario está unido y el tema de la descripción.

Además, se realizó un estudio de los usuarios a partir de sus tweets, donde se investigaron patrones en los temas de los tweets de odio, la diversidad de los objetivos de discriminación de los haters, la incendiabilidad de los tweets de odio y la relación entre la personalidad y el tipo de usuario. Se encontró que la mayoría de los usuarios se centran en discriminar a un único colectivo, y que en general, los tweets de odio son menos incendiarios que los de no odio. Sin embargo, no se encontró una diferencia significativa en la personalidad entre los diferentes tipos de usuarios.

El objetivo de la última parte de este proyecto es analizar las relaciones entre los diferentes tipos de usuarios en una red social. Se busca comprender si aquellos que responden a mensajes ofensivos también son emisores de odio, simples observadores o se oponen al discurso de odio. Este objetivo se desglosa en dos partes principales: primero, identificar con qué tipo de usuarios suelen relacionarse los usuarios de cada tipo (hater, upstander y neutro); y segundo, determinar la naturaleza de estas interacciones, es decir, si son de apoyo, indiferencia u oposición.

Para llevar a cabo este análisis se utiliza el diseño de un grafo, donde cada usuario se representa como un nodo y las interacciones entre ellos se muestran como aristas dirigidas. El color de los nodos indica el tipo de usuario (hater, neutro, upstander), mientras que el color de las aristas refleja la naturaleza de la interacción (apoyo, indiferencia, oposición).

La obtención de las aristas implica filtrar los tweets extraídos, traducirlos y procesarlos con una inteligencia artificial que determina el tipo de relación argumentativa entre ellos. Posteriormente, se clasifican todos los usuarios en hater, upstander o neutro utilizando un clasificador basado en las características de sus perfiles.

Los resultados muestran que la mayoría de las interacciones son de apoyo y que la mayoría de los usuarios son identificados como haters. Esto sugiere una tendencia de los usuarios a relacionarse con aquellos que comparten sus mismas opiniones y actitudes. Los grafos generados visualizan claramente que los haters tienden a interactuar principalmente entre sí.



En conclusión, este análisis revela la formación de comunidades virtuales basadas en el odio, donde los usuarios se comunican preferentemente con individuos de mentalidad similar. Cada grafo representa un grupo de usuarios que expresan odio hacia un colectivo específico, lo que destaca la importancia de abordar de manera integral el fenómeno del discurso de odio en línea.

En resumen, este proyecto ha demostrado que es posible clasificar usuarios de Twitter en función de su perfil con una precisión considerable, lo que puede ser útil para identificar discursos de odio en la plataforma y tomar medidas adecuadas para combatirlos.



Anexo 1

Manual de uso del material entregado

1. Clasificación del usuario en función del perfil

Dentro de la carpeta de Drive proporcionada, titulada “Material Proyecto REAL-UP”, hay una subcarpeta llamada “Notebooks”. En esta subcarpeta se encuentra otra denominada “Clasificador de usuarios”. Aquí se halla el material diseñado para clasificar a un usuario según su perfil. Este recurso ha sido elaborado específicamente para ser utilizado por personas que no necesariamente posean conocimientos en Python, siendo la opción más sencilla e intuitiva. Contiene tres archivos:

1. `clasificador-usuarios.ipynb`: El cuaderno de Python con el que se interactúa para clasificar a un usuario.
2. `bancodenumbers.zip`: Contiene cuatro ficheros necesarios para identificar el nivel de anonimato y el género de un usuario. Es necesario descargarlo, descomprimirlo y ponerlo en la misma carpeta que el cuaderno de Python.
3. `model.json`: El árbol de decisión que clasifica a un usuario dadas las características de su perfil. Hace falta descargarlo e incluirlo en la misma carpeta que el cuaderno de Python.

Tras haber descargado todos estos archivos, para clasificar a un autor es necesario abrir el cuaderno de Python. En él, es necesario ejecutar todas las celdas, cambiando los valores de ejemplo por los del perfil del usuario que se quiera clasificar. Para conseguir estas características no es necesario usar la API de Twitter, sino que se pueden extraer directamente desde la interfaz de Twitter en la aplicación del móvil (si el usuario es público).

2. Datasets

En la carpeta de Drive entregada, con nombre “Material Proyecto REAL-UP”, se encuentra una subcarpeta, “Datasets” que contiene dos subcarpetas:

1. “Tweets extraídos en Julio y Agosto 2023”: la colección de tweets extraídos a partir de las keywords del discurso de odio. Si se descomprime, se verá que el formato de cada archivo es `oberaxe_<tipo de keyword>_<fecha>`, terminando en `replies` o `no`, dependiendo de si son las respuestas o no.

2. “Demás archivos generados”: la colección de archivos generados a partir del análisis descrito en este informe.

3. Notebooks desarrollados para el análisis

En la carpeta de Drive entregada, con nombre “Material Proyecto REAL-UP”, se encuentra una subcarpeta, “Notebooks”; dentro de ella, la subcarpeta “Notebooks utilizados para el análisis”. Contiene todos los cuadernos de Python que se usaron para este trabajo. Se incluyen para que el análisis pueda ser realizado en el futuro con nuevos tweets, puesto que al final este proyecto se ha realizado sobre una foto estática de Twitter.

1. `analisis.ipynb`: Contiene el código que se usó para realizar un primer análisis de los tweets extraídos en julio y agosto y para escoger a los usuarios prolíficos. La última parte de este archivo incluye tests de clustering con la librería `fuzzy` que más tarde no se usaron para la clasificación definitiva.
2. `extraccion_caracteristicas.ipynb`: Contiene el código que finalmente se usó para el clustering de usuarios y la siguiente extracción de características del perfil.
3. `entrenamiento.ipynb`: Contiene el código de entrenamiento del clasificador. Este archivo se ejecutó en Google Colab, estando conectado a una GPU NVIDIA Tesla T4.
4. `analisis2.ipynb`: Contiene el código usado para extraer los resultados del [Capítulo 7](#). Este archivo usa Spark para paralelizar el procesamiento de los tweets y la GPU para acelerar la clasificación con varios modelos. Si no se dispone de un clúster de ordenadores o de una GPU, respectivamente, entonces haría falta modificar el código para no incorporar estas herramientas.
5. `grafo_prueba.ipynb`: Contiene el código que en un principio se iba a usar para la creación del grafo. En particular, incluye el grafo de la conversación más larga de entre todas las conversaciones con keywords relacionadas con las personas gitanas. Aunque al final no se usara, se incluye aquí por si es de interés para futuros trabajos.
6. `grafo.ipynb`: Contiene el código que se usó para la extracción de las aristas y nodos descrita en el [Capítulo 8](#). La última [sección, “2.º Grafo”](#), incluye un código inicial en el que se usaba la librería `networkx` para crear el grafo, en vez de `Visone`. Esto tampoco llegó a entrar en el proyecto, pero se deja por si es de interés.
7. `grafo_personas_conocidas.ipynb`: Incluye un análisis adicional que no formaba parte de este proyecto, sino que fue realizado de manera independiente debido a su interés, y quizás pueda ser una dirección futura de investigación. Se describe en el [Anexo C](#).

4. Grafos de conversaciones

En la carpeta “Grafos de conversaciones” se hallan dos subcarpetas, “Imágenes” y “Grafos Interactivos”. La primera contiene seis imágenes, cada una correspondiente al grafo de las conversaciones sobre un tipo de colectivo (personas gitanas, judías, migrantes, minorías culturales y religiosas, musulmanas y general). La segunda incluye los archivos desde los cuales se han exportado las imágenes anteriores. La extensión de éstos es `.graphmlz`, un formato de archivo comprimido para grafos. Para abrirlos, es recomendable usar el programa **Visone**, que no solo permite visualizar el grafo, sino explorarlo y analizarlo.

4.1. Instalación de Visone

El software Visone está escrito en Java, lo que significa que para ejecutarlo en un ordenador, es necesario tener Java instalado. La propia aplicación se descarga desde [la página de descargas de Visone](#). En la sección **current version** aparecen las versiones más recientes.

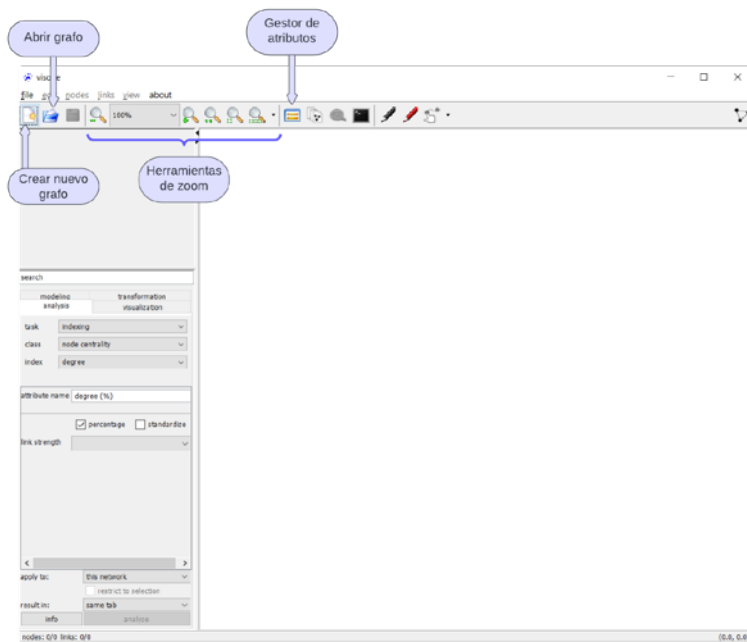
La primera es `visone-2.27.1.jar`⁸, apta para todos los sistemas operativos, y la segunda es `visone-2.27.1-win64.zip`, apta únicamente para Windows de 64 bits. La ventaja de esta segunda es que ya integra un entorno de ejecución de Java, mientras que para usar la primera es necesario tener Java instalado⁹.

Tras haber descargado el archivo de la página de Visone, se descomprime, se accede a la carpeta descomprimida y se hace doble clic en la aplicación de Visone. Cuando se ejecuta el programa, cuya versión actual está en inglés, se abre la pantalla principal ([Figura 29](#)).

8 A día 26/01/24.

9 Se puede instalar a través de la página [de descargas de Java](#), en la sección JRE 8.

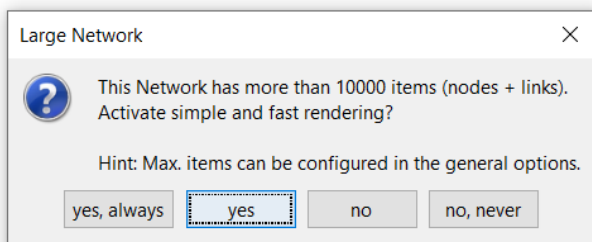
Figura 29: Página de inicio de Visone. En el menú superior se sitúan las tres herramientas principales.



4.2. Visualización de grafos

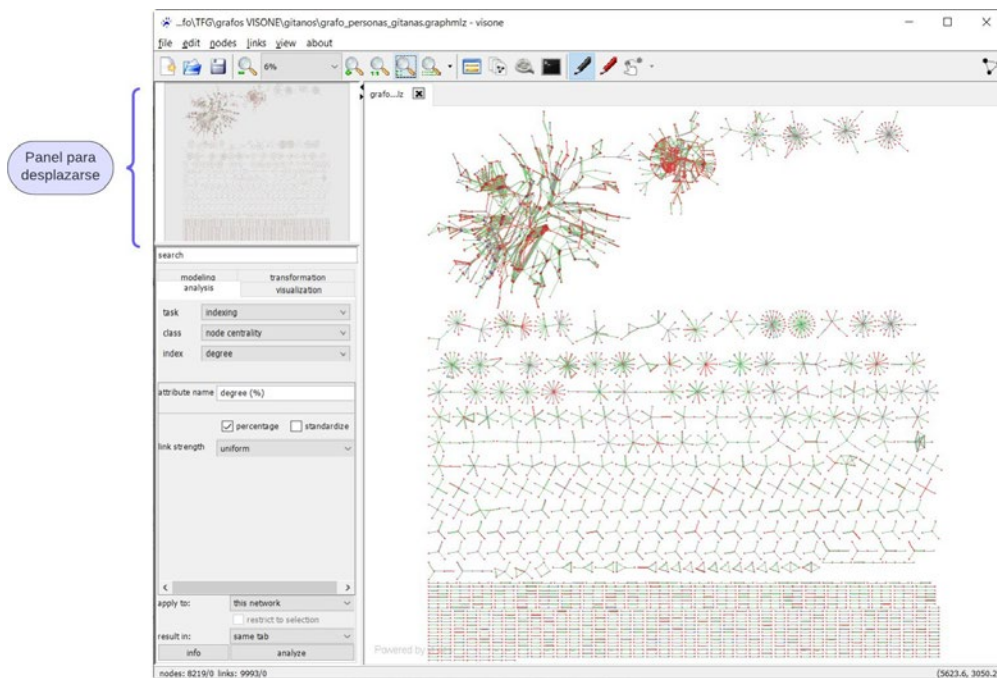
Para ilustrar lo siguiente, digamos que queremos visualizar la red formada por las conversaciones sobre el colectivo gitano. Entonces, es necesario haber descargado el archivo `user_edges_personas_gitanas.graphmlz` de la carpeta de Drive. Para abrirlo en Visone, pulsamos el icono azul de Abrir Grafo (Figura 29). Aparece un aviso que nos informa de que es una red de gran tamaño y pregunta cómo queremos tratarla (Figura 30).

Figura 30: Aviso de red de gran tamaño. Pulsar “yes”.



Tras pulsar “yes”, se visualiza el grafo mostrado a continuación (Figura 31).

Figura 31: Grafo de conversaciones sobre el colectivo gitano en el programa Visone.



Para desplazarse por el grafo se puede usar el panel superior izquierdo (Figura 31) o haciendo clic derecho. Para agrandar la imagen, se pueden usar las herramientas de zoom (Figura 30) o la rueda del ratón. El significado de los colores de los nodos y de las conexiones se explica en la [Sección 8.1](#).

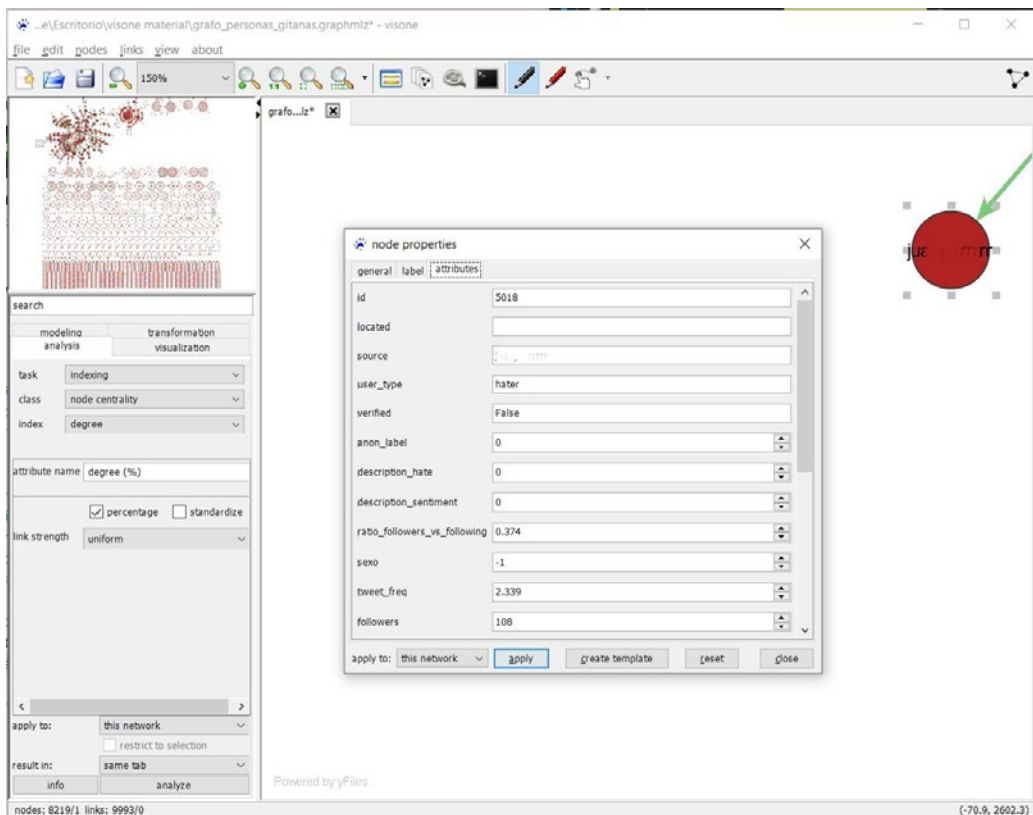
4.3. tributos de un nodo

Quizá sea interesante conocer las características del perfil que han sido usadas para clasificar a un usuario concreto. Estas se pueden ver de la siguiente manera:

1. Pinchar en un nodo
2. En el menú superior (donde aparecen las opciones file, edit, nodes, links, ...), pinchar en nodes.
3. Pinchar en la última opción, Properties.
4. Aparece una ventana con tres pestañas: general, label y attributes. Marcar attributes.

Si se han seguido estos pasos, la pantalla se parecerá a la [Figura 32](#). Las características y sus valores están descritos en la sección ...

Figura 32: Detalle de un nodo: las características del perfil que han sido usadas para clasificar al usuario @userAnonimo.



4.4. Explorando el grafo: importancia de cada usuario

Dentro de un grafo, cada nodo tiene asociado varias medidas de “importancia” que resumen diferentes aspectos de su papel dentro de la red. En términos generales, cuando más grande sea el valor de estas medidas, más conectado está el usuario en la red, y por ende, más relevancia tiene dentro de su entorno. En el caso de redes simples, basta leer el valor numérico de estas medidas para hacerse una idea del peso individual de cada nodo, pero cuando una red tiene muchos nodos y aristas, esta tarea resulta laboriosa. Es por ello que, en redes complejas, conviene representar el tamaño de los nodos en función de su importancia.

Por ejemplo, imaginemos que queremos visualizar los nodos del subgrafo que aparece en la esquina superior izquierda de la [Figura 31](#), que está compuesto por múltiples nodos y aristas que se superponen. Para ello, primero lo seleccionamos y hacemos Ctrl+C para copiarlo. Después, creamos un nuevo grafo (con el icono indicado en la [Figura 29](#)) y hacemos Ctrl+V para pegarlo en la pestaña que se acaba de abrir.

Figura 33: Selección del subgrafo complejo.

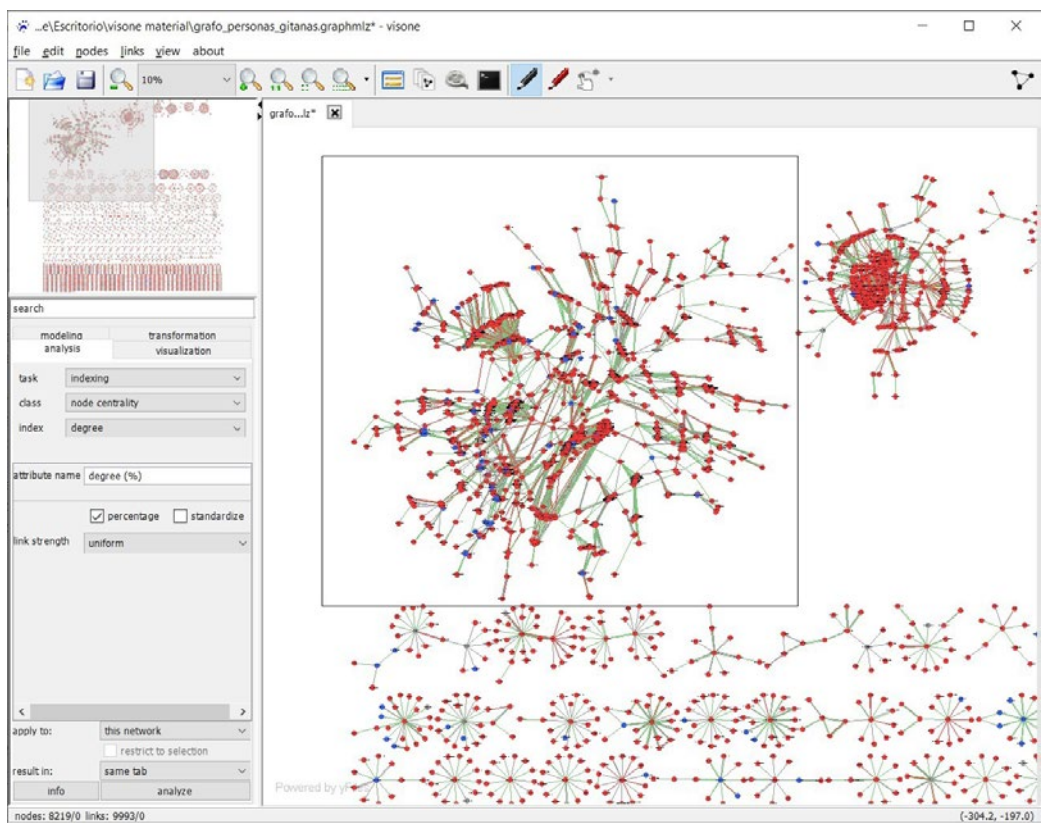
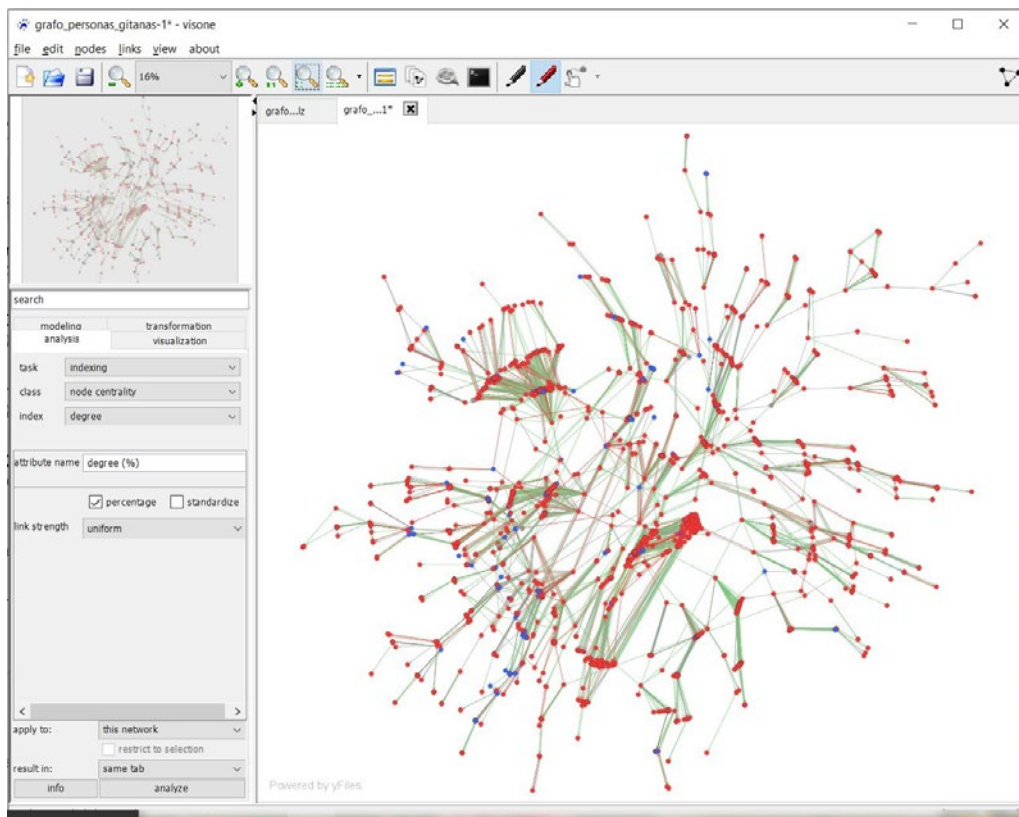
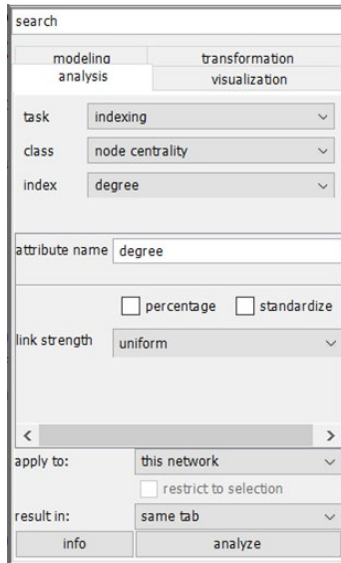


Figura 34: Al abrir una pestaña nueva, pegamos el grafo anteriormente seleccionado.



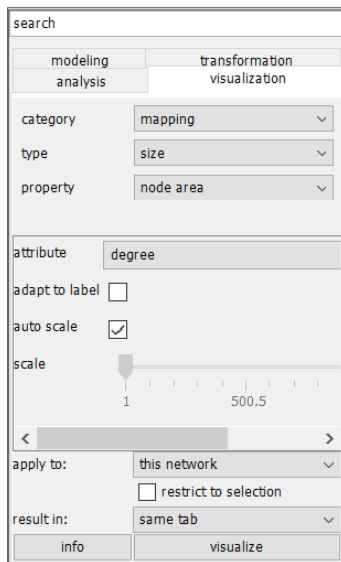
Para representar cada nodo en función de su importancia dentro de la red, se deben seguir los pasos descritos a continuación.

Figura 35: Opciones en el panel de Visone requeridas para el cálculo del *degree*, una medida de centralidad de los nodos.



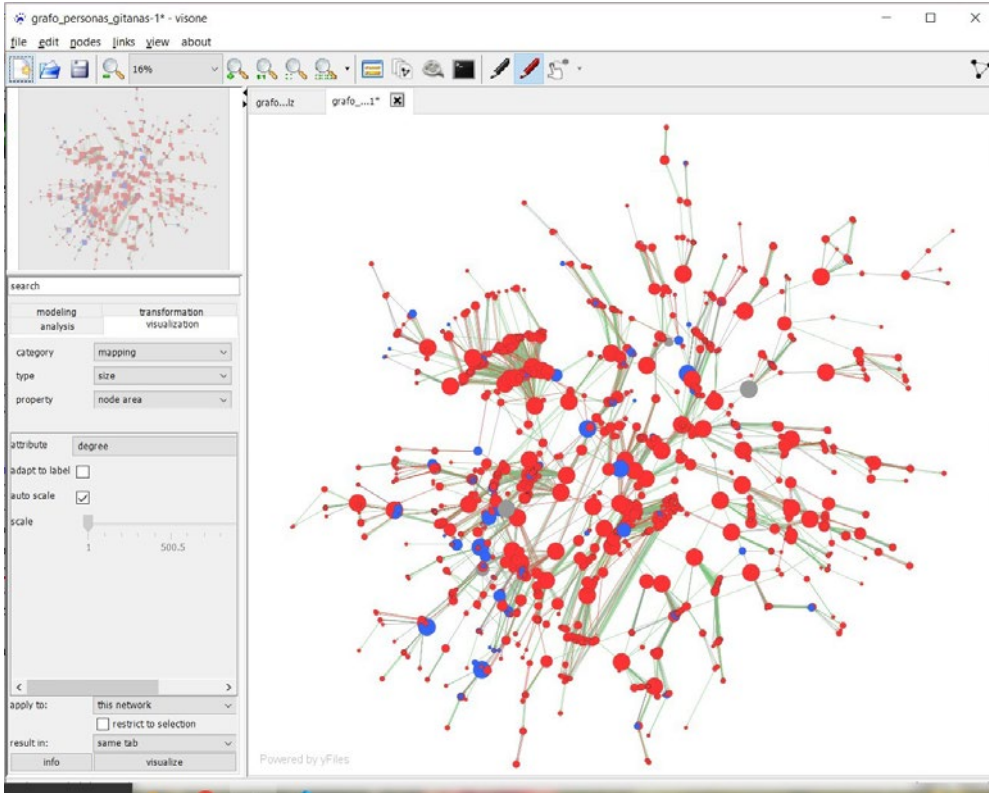
1. En el panel izquierdo seleccionamos la pestaña analysis.
2. Marcamos las opciones indexing en task y node centrality en class.
3. Elegimos la medida que queramos en index (degree, indegree, betweenness, ...), por ejemplo, degree.
4. Desmarcamos la opción percentage para obtener el valor real del grado.
5. Finalizamos con el botón analyze. Tras seguir estos pasos, cada nodo ahora dispone de un nuevo atributo: el valor de la medida calculada.

Figura 36: Opciones en el panel de Visone para la visualización del *degree* de cada nodo.



1. En el panel izquierdo seleccionamos la pestaña visualize.
2. Marcamos las opciones mapping en category, size en type y node area en property.
3. En attribute, seleccionamos el atributo que acabamos de calcular: *degree* (o el nombre correspondiente a la medida que se haya usado).
4. Finalizamos con el botón visualize. Si todo se ha realizado correctamente, ahora Visone mostrará el grafo de la [Figura 37](#).

Figura 37: Subgrafo complejo, donde el tamaño de cada nodo representa su importancia dentro de la red (según la medida de centralidad calculada, en este caso *degree*).



En conclusión, ahora podemos observar la red de manera más detallada y extraer conclusiones sobre si los usuarios odiadores están más conectados o menos que los otros tipos de usuarios. En este ejemplo, podríamos deducir que tanto los haters como los upstanders son importantes en la red, pero que, de entre los poco relevantes, hay más haters.

4.5. Explorando el grafo: filtrando nodos por un atributo específico

Finalmente, es interesante analizar, por ejemplo, cómo se relacionan los usuarios upstander entre sí, sin tener de por medio a los demás tipos de usuarios. El programa Visone permite filtrar los nodos por su atributo `user_type`, de la siguiente manera.

1. Desde el grafo de la [Figura 34](#), pinchamos en el gestor de atributos (indicado en la [Figura 29](#)).
2. Elegimos las opciones mostradas en la [Figura 38](#) para seleccionar a los nodos haters y neutros. En el menú superior, pinchamos en edit → cut.

Figura 38: Selección de los nodos correspondientes a haters y neutros.

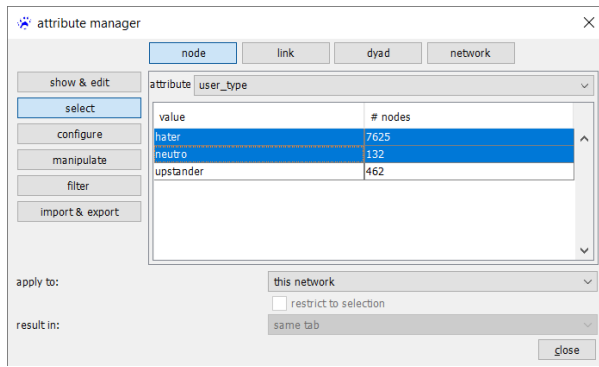
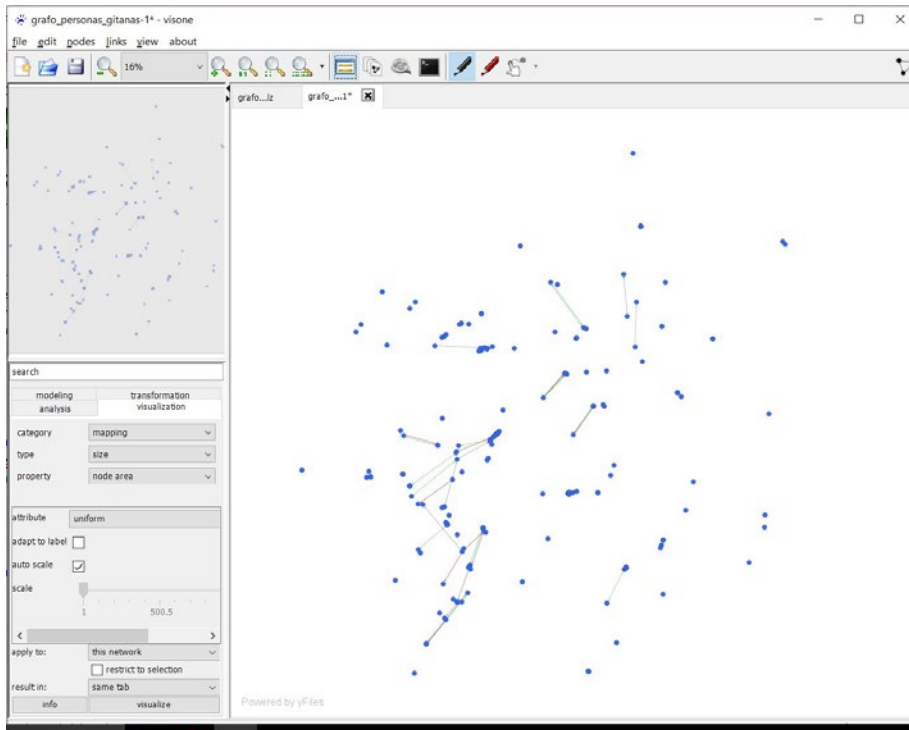


Figura 39: Visualización únicamente de los nodos upstander y sus conexiones.



En este caso, se observa que los upstander se relacionan entre sí de todas las maneras posibles puesto que hay conexiones rojas (oposición), verdes (apoyo) y grises (neutro).

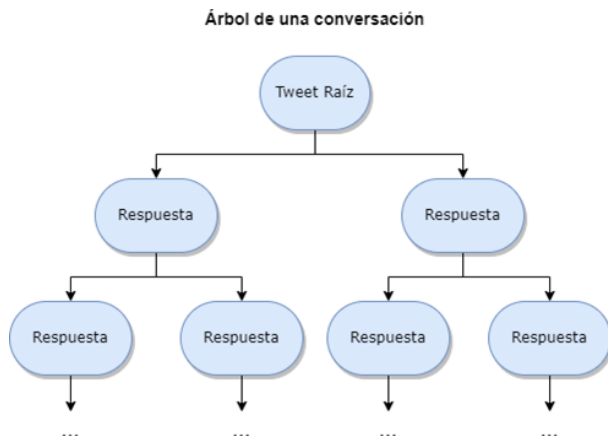
Anexo 2

Sobre conversaciones en Twitter

Para analizar las interacciones entre usuarios en Twitter, es esencial descargar las conversaciones asociadas con cada tweet. De acuerdo con la definición de Twitter, una conversación se compone de todas las respuestas generadas a partir de una publicación. Esto incluye las respuestas directas, las respuestas a esas respuestas y así sucesivamente.

Para rastrear internamente a qué conversación pertenece un tweet, Twitter asigna a cada uno un identificador único que coincide con el identificador del tweet que la generó. Tal como está implementado en Twitter, los tweets que se crean desde cero o que citan a otro tweet son los tweets “raíz” inician una conversación. Los demás tipos de tweets (retweets y respuestas a un tweet) no inician nuevas conversaciones, sino que forman parte de una conversación existente con un identificador diferente al suyo. De ahora en adelante, para simplificar el texto, llamaremos “tweets raíz” a los tweets que generan conversaciones y “tweets derivados” a los que no lo hacen. La [Figura 40](#) muestra el gráfico de una conversación.

Figura 40: Diagrama del método de extracción de tweets.



Es posible obtener todos los mensajes de una conversación a través de una consulta simple a la API, usando el identificador del tweet raíz. Sin embargo, lo que no permite la API es extraer todos los mensajes generados a partir de un tweet derivado puesto que el identificador de la conversación no es el mismo que

el del tweet derivado. Es cierto que se podría extraer toda la conversación de la que es parte, en ella se encontrarían las respuestas al tweet que nos interesa, pero esta opción se ha descartado porque significa recorrer todos los tweets de una conversación mucho más extensa, la cual no nos concierne, a cambio de un tiempo de ejecución mucho mayor. No obstante, sí que podemos extraer las respuestas directas de estos tweets derivados. Todo esto se explica de manera gráfica en las figuras 41 y 42.

Figura 41: Cuando el tweet cuya conversación nos interesa conocer es raíz, entonces podemos extraer todas sus respuestas derivadas.

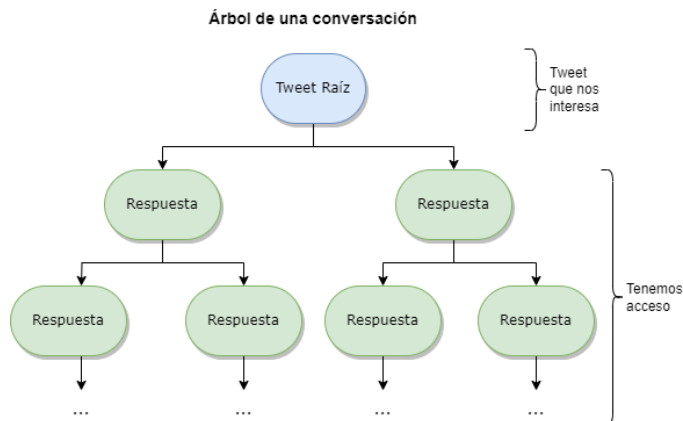
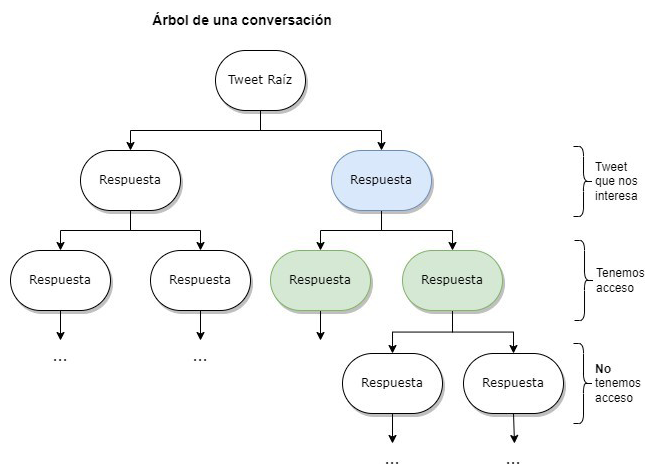


Figura 42: Cuando el tweet cuya conversación nos interesa conocer es derivado, entonces solo podemos extraer sus respuestas directas.



Nótese que el hecho de que las citas se consideren como tweets raíz implica que no forman parte de la conversación del tweet al que citan. Esto en algunos casos va a significar que nos quedaremos sin información parcial, pues las citas, en su concepto más simple, son también respuestas a un tweet.

Anexo 3

Grafo de personas conocidas

Como análisis adicional que no formaba parte de este proyecto, se realizó un estudio de manera independiente debido a su interés. Quizás pueda ser una dirección futura de investigación.

Cada tweet que se extrajo por keyword tiene asociados múltiples atributos (fecha, número de retweets, número de respuestas, etc.). En particular, uno de estos atributos es una colección de lo que en Twitter se denomina “context annotations” (anotaciones del contexto), formada por entidades reconocidas que Twitter reconoce en el texto del tweet. Estas entidades pueden ser películas, eventos deportivos, noticias de actualidad... Sin embargo, nos interesan unas en concreto: las entidades que se refieren a personalidades célebres. Twitter usa varias categorías para diferentes tipos de personalidades:

- *Persona, Político, Músico, Actor, Atleta, Personalidad del deporte, Entrenador, Periodista, Creador digital, Personalidad de los videojuegos.*

El objetivo es analizar cómo los usuarios, en particular los odiadores, cuando hablan de estas personalidades conocidas, cómo se posicionan. Si es con connotación negativa, neutra o positiva.

Para ello, primero seleccionamos un tipo de odio, el que ataca a las personas gitanas (usamos los archivos con keywords relacionados con personas gitanas). Cargamos los tweets que contengan a personas célebres y los traducimos con el modelo de [Hug-gingFace Helsinki-NLP/opus-mt-es-en](#). A continuación, los clasificamos con el modelo [eevvgg/StanceBERTa](#), que indica si el posicionamiento del autor del texto hacia la entidad mencionada es positiva, negativa o neutral.

El resultado es el archivo cuyas filas iniciales se muestran en la tabla siguiente.

Tabla 13: Archivo `user_edges_entities.csv`. La primera columna son los valores internos que usa Twitter para identificar a cada usuario; la segunda, las personas conocidas; la tercera, de qué manera se posicionó el usuario para hablar de la persona conocida en un tweet.

source	target	relation
1683260011906998273	Santiago Abascal	negative
1683251339969671168	Pedro Sánchez	neutral
1683251084352012288	Pedro Sánchez	negative
1683250494909579264	Pedro Sánchez	negative
1683249504160567302	Álvaro Uribe	negative
1683248993239695360	Santiago Abascal	negative
1683248473900982273	Santiago Abascal	negative
1683247497697714177	Alberto Núñez Feijóo	neutral
1683238278135259143	Pedro Sánchez	neutral
1683235984765251585	Vladimir Putin	neutral
1683232881076756481	Pedro Sánchez	neutral
1683232102702018561	Ben Shapiro	negative
1683225378154749962	Carlos Rivera	positive
1683225378154749962	Carlos Rivera	positive
.	.	.

Si se desea continuar con esta línea de trabajo, se puede usar Visone para visualizar el grafo que se genera a partir de este archivo. Además, si se quiere ejecutar el análisis para otros tipos de odio, el cuaderno de Python que contiene el código correspondiente es `grafo_personas_conocidas.ipynb`, en la carpeta “Material Proyecto REAL-UP” → “Notebooks” → “Notebooks utilizados para el análisis”.

Bibliografía

- [1] *Twitter replaces its free API with a paid tier in quest to make more money.* <https://www.theverge.com/2023/2/2/23582615/twitter-removing-free-api-developer-apps-price-announcement>. [Online; accessed 22/10/2023] (vid. pág. 7).
- [2] Gloria del Valle-Cano, Lara Quijano-Sánchez, Federico Liberatore y Jesús Gómez. «SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles». En: *Expert Systems with Applications* 216 (2023), doi: [10.1016/j.eswa.2022.119446](https://doi.org/10.1016/j.eswa.2022.119446) (vid. págs. 9, 13, 37).
- [3] Kai Shu, Suhang Wang y Huan Liu. «Understanding User Profiles on Social Media for Fake News Detection». En: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. Abr. de 2018, págs. 430-435. doi: [10.1109/MIPR.2018.00092](https://doi.org/10.1109/MIPR.2018.00092) (vid. pág. 12).
- [4] A. Pastor López-Monroy, Fabio A. González y Thamar Solorio. «Early author profiling on Twitter using profile features with multi-resolution». En: *Expert Systems with Applications* 140 (2020), pág. 112909. issn: 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2019.112909>. url: <https://www.sciencedirect.com/science/article/pii/S095741741930627X> (vid. pág. 12).
- [5] Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara et al. «TweetNLP: Cutting-Edge Natural Language Processing for Social Media». En: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Abu Dhabi, U.A.E.: Association for Computational Linguistics, nov. de 2022 (vid. págs. 13, 34, 37).
- [6] *Github repository of TweetNLP Project.* <https://github.com/cardiffnlp/tweetnlp> (vid. pág. 13).
- [7] Sai Teja Peddinti, Keith W. Ross y Justin Cappos. «„On the Internet, Nobody Knows You’re a Dog”: A Twitter Case Study of Anonymity in Social Networks». En: *Proceedings of the Second ACM Conference on Online Social Networks*. COSN '14. Dublin, Ireland: Association for Computing Machinery, 2014, págs. 83-94. isbn: 9781450331982. doi: [10.1145/2660460.2660467](https://doi.org/10.1145/2660460.2660467). url: <https://doi.org/10.1145/2660460.2660467> (vid. pág. 26).
- [8] Twitter. *X Verification requirements how to get the blue check.* <https://help.twitter.com/en/managing-your-account/about-x-verified-accounts>. [Online; accessed 21/10/2023] (vid. pág. 26).
- [9] *GitHub repository of Marc Boquet.* <https://github.com/marcboquet/spanishnames> (vid. págs. 26, 29).

- [10] *Github repository of Christos Baziotis*. <https://github.com/cbaziotis/ekphrasis> (vid. pág. 26).
- [11] Robert R McCrae y Paul T Costa. «Validation of the five-factor model of personality across instruments and observers.» En: *Journal of personality and social psychology* 52.1 (1987), pág. 81 (vid. pág. 29).
- [12] Danny Azucar, Davide Marengo y Michele Settanni. «Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis». En: *Personality and Individual Differences* 124 (2018), págs. 150-159. issn: 0191-8869. doi: <https://doi.org/10.1016/j.paid.2017.12.018>. url: <https://www.sciencedirect.com/science/article/pii/S0191886917307328> (vid. pág. 29).

